85 Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie

JANNIK WILHELM

Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage



Jannik Wilhelm

Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage

Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie (KIT)
Band 85

Herausgeber: Prof. Dr. C. Hoose
Prof. Dr. P. Knippertz
Prof. Dr. J. G. Pinto

Institut für Meteorologie und Klimaforschung
am Karlsruher Institut für Technologie (KIT)
Kaiserstr. 12, 76128 Karlsruhe

Eine Übersicht aller bisher in dieser Schriftenreihe
erschienenen Bände finden Sie am Ende des Buches.

Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage

von
Jannik Wilhelm



Karlsruher Institut für Technologie
Institut für Meteorologie und Klimaforschung

Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften von der KIT-Fakultät für Physik des Karlsruher Instituts für Technologie (KIT) genehmigte Dissertation

von Jannik Wilhelm, M.Sc.

Tag der mündlichen Prüfung: 12. Februar 2021
Referent: Prof. Dr. Michael Kunz
Erster Korreferent: Prof. Dr. Roland Potthast
(University of Reading, Deutscher Wetterdienst)
Zweiter Korreferent: Prof. Dr. Christoph Kottmeier

Impressum



Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

www.ksp.kit.edu





Print on Demand 2022 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 0179-5619
ISBN 978-3-7315-1182-3
DOI 10.5445/KSP/1000145542

Kurzfassung

Die vorliegende Dissertation beschäftigt sich mit der Analyse der Lebenszyklen konvektiver Zellen im Zusammenhang mit den vorherrschenden Umgebungsbedingungen in Deutschland. Darüber hinaus werden verschiedene statistische Vorhersagemodelle zur Abschätzung der Lebensdauer und der Größe konvektiver Zellen entwickelt und untersucht. Das Ziel dabei ist es herauszufinden, welche Methode für eine Verbesserung von Verfahren zur Echtzeit-Vorhersage (*Nowcasting*) am besten geeignet ist.
Die Grundlage für diese Untersuchungen bilden Daten des radarbasierten Zellverfolgungsalgorithmus KONRAD, anhand derer zusammenhängende Lebenszyklen von isolierter Konvektion (Einzel- und Superzellen) für die Sommerhalbjahre 2011 – 2016 erstellt werden. Zusätzlich wird eine Vielzahl konvektionsrelevanter Umgebungsvariablen unter Verwendung von hochaufgelösten Assimilationsanalysen des numerischen Wettervorhersagemodells COSMO-EU berechnet und mit den Lebenszyklen zusammengeführt. Auf Basis dieses kombinierten Datensatzes werden statistische Zusammenhänge zwischen verschiedenen Zellattributen und Umgebungsvariablen untersucht. Wie die Analysen zeigen, sind insbesondere Maße der vertikalen Windscherung aufgrund ihres Einflusses auf die Organisationsform der Konvektion geeignet, zwischen Zellen mit kurzer und langer Lebensdauer zu unterscheiden. Eine erhöhte thermische Instabilität der Atmosphäre geht mit einem schnelleren anfänglichen Wachstum der Zellen einher, welches wiederum eine größere horizontale Zellausdehnung (Zellfläche) während des Lebenszyklus und damit indirekt eine längere Lebensdauer begünstigt.

Drei unterschiedliche multivariate Methoden (logistische Regression, *Random Forest*, nicht-linearer Polynomansatz) werden als Modelle für die Abschätzung der Lebensdauer und der maximalen Zellfläche der konvektiven Zellen mit Hilfe eines Ensembleansatzes untersucht. Die Vorhersagegüte der Modelle wird evaluiert und die Bedeutung der anfänglichen Zellentwicklung und der Umgebungsvariablen analysiert. Dabei werden Potentiale und Grenzen der Methoden aufgezeigt, die verdeutlichen, dass die Wahl eines geeigneten Verfahrens von der genauen Fragestellung bzw. Anforderung des *Nowcastings* abhängt. Die Untersuchungen legen dar, dass sich die maximale Zellfläche der konvektiven Zellen insgesamt besser abschätzen lässt als ihre Lebensdauer. Umgebungsvariablen, die den dynamischen und thermodynamischen Zustand der Atmosphäre charakterisieren, sind insbesondere zu Beginn der Zellentwicklung für die Abschätzung der zukünftig zu erwartenden Entwicklung der Zellen bedeutsam, während mit zunehmendem Zellalter die vergangene Zellhistorie immer wichtiger wird.

Inhaltsverzeichnis

1 Einleitung und wissenschaftliche Fragestellungen

Gewitter zählen in vielen Teilen der Erde zu den bedeutsamen Wettererscheinungen. Als Folge hochreichender Feuchtkonvektion in der Atmosphäre beeindrucken sie nicht nur durch ihr imposantes Erscheinungsbild, sondern bergen gleichzeitig ein hohes Gefahren- und Schadenpotential. Im Vergleich zu der allgemeinen Zirkulation der Atmosphäre und dem gesamten Wasserkreislauf des Klimasystems transportieren Gewitter als ein Teil dieser Systeme zwar deutlich geringere Mengen an Energie und Feuchte (z. B. Israël, 1961; Kraus, 2004; Vallis, 2017), können jedoch auf kurzen Zeitskalen von einer bis wenigen Stunden lokal sehr große Energiemengen durch Phasenumwandlungen des in der Atmosphäre befindlichen Wasserdampfs freisetzen. Diese können wiederum verschiedene meteorologische Phänomene hervorrufen. Neben Blitzentladungen können Gewitter mit weiteren für Mensch, Tier, Eigentum, Infrastruktur und (Land-)Wirtschaft gefährlichen und schadenträchtigen Begleiterscheinungen wie Starkregen, Hagel, Starkwindböen und Tornados einhergehen.
Generell bedarf es zur Entstehung von hochreichender Konvektion geeigneter Voraussetzungen, die von vielen Prozessen auf unterschiedlichen Raum- und Zeitskalen abhängen. Neben einem ausreichenden Feuchteangebot in der unteren Troposphäre sind eine labile Schichtung und ein Mechanismus erforderlich, der vertikal ausgelenkten Luftpaketen einen freien Aufstieg durch thermischen Auftrieb ermöglicht (Doswell, 1987; Johns und Doswell, 1992). Auf der lokalen bzw. der Mesoskala geschieht dies mittels Hebung, die

durch unterschiedliche Mechanismen ausgelöst werden kann wie beispielsweise durch horizontale Strömungskonvergenzen, thermische Windsysteme in orografisch gegliedertem Gelände, atmosphärische Schwerewellen oder Querzirkulationen an synoptisch-skaligen Konvergenzlinien und Fronten (z. B. Markowski und Richardson, 2010). Großräumige Hebung durch synoptisch-skalige Wettersysteme hingegen führt durch adiabatische Abkühlung zu einer großflächigen Destabilisierung und Anreicherung von Feuchte in der unteren und mittleren Troposphäre, die den freien Aufstieg von Luftpaketen erleichtert (Trapp, 2013). Erreicht die Luft eines aufsteigenden Luftpakets Sättigung, kommt es zur Wolkenbildung. Ab dem Niveau freier Konvektion erfährt das Luftpaket durch thermischen Auftrieb eine vertikale Beschleunigung, welche zur Ausbildung eines sich vertikal intensivierenden Aufwindbereichs führt, sodass Cumulonimbuswolken mit einer vertikalen Mächtigkeit von etwa 10 (mittlere Breiten) bis 16 km (Tropen) entstehen können. Erst am Niveau des neutralen Aufstiegs endet die freie Konvektion. Dieses Niveau liegt in der Regel in der Höhe der Tropopause, welche durch eine Temperaturinversion gekennzeichnet ist und daher einen weiteren thermischen Auftrieb unterbindet. Im Fall besonders starker Aufwinde kann die Luft aufgrund ihrer Trägheit auch geringfügig in die untere Stratosphäre eindringen und beispielsweise im Satellitenbild als konvektives Überschießen (*Overshooting Top*) beobachtet werden. Mit Beginn des fallenden Niederschlags entstehen Abwindbereiche, die sehr große negative Vertikalgeschwindigkeiten erreichen können. Aufgrund dieser dynamischen Strukturierung spricht man von der Entstehung einer konvektiven Zelle.

In Mitteleuropa kommt es insbesondere im Sommerhalbjahr von April bis September zu teils starken konvektiven Ereignissen (Taszarek et al., 2019). Die meisten von ihnen gehen als Gewitter mit Blitzen einher, die elektrische Entladungen nach einer Ladungstrennung in den Wolken darstellen, welche sich durch Kollisionen der Hydrometeore sowie induktive Prozesse einstellt. Eine häufige Begleiterscheinung konvektiver Zellen ist Starkregen. Je stärker

und breiter der Aufwindbereich in einer konvektiven Zelle ist, desto effizienter laufen auch die Kondensation und die niederschlagsbildenden Prozesse ab, da im Kernbereich des Aufwinds das Einmischen trockener Umgebungsluft eine geringere Rolle spielt (z. B. Doswell et al., 1996; Trapp, 2013). Besonders langsam ziehende konvektive Zellen oder eine Sequenz mehrerer aufeinander folgender Zellen können zu hohen akkumulierten Niederschlagssummen und hohen Abflussraten in sehr kurzer Zeit führen. Lokale Überschwemmungen und – bei entsprechender topografischer Geländestrukturierung – Sturzfluten können die Folge sein. In Deutschland kam es beispielsweise während einer zweiwöchigen Periode von Ende Mai bis Anfang Juni 2016 besonders in der Südhälfte des Lands zu zahlreichen Gewitterereignissen, die zu teils schweren Überschwemmungen führten (Piper et al., 2016). Sturzfluten in Braunsbach sowie in Simbach am Inn zerstörten ganze Straßenzüge (Bronstert et al., 2017; Hübl, 2017; Vogel et al., 2017). In dieser Periode starben insgesamt elf Menschen und es entstand ein Gesamtschaden von etwa 2,6 Milliarden Euro (versicherte Schäden 1,2 Milliarden Euro; Munich Re, 2017).

Eine weitere Begleiterscheinung konvektiver Zellen ist Hagel. Dieser entsteht vornehmlich in langlebigen konvektiven Zellen mit einem starken Aufwindbereich. Ist eine hohe Konzentration von unterkühlten Wolkentröpfchen und Eispartikeln in der Wolke vorhanden, setzt Hagelbildung ein (Pruppacher und Klett, 2010). Je nach Verweildauer der entstehenden Hagelkörner erreichen diese unterschiedlich große Durchmesser zwischen 0,5 und mehr als 10 cm. In Deutschland fällt Hagel an einem Ort an etwa null bis vier Tagen pro Jahr, wobei die höchsten Werte in den Bereichen der

Schwäbischen Alb, des Alpenvorlands, des Erzgebirges und des Rhein-Main-Gebiets beobachtet werden (Puskeiler et al., 2016; Schmidberger, 2018). Erreichen die Hagelkörner große Durchmesser von mehreren Zentimetern, können Hagelunwetter hohe Schadensummen verursachen, wenn sie in den betroffenen Gebieten viele vulnerable Objekte wie Gebäude, Fahrzeuge, Infrastrukturen oder landwirtschaftliche Erzeugnisse beschädigen. Beispielsweise verursachten Superzellen am 27. Juli 2013 in einer Region um Hannover und am 28. Juli 2013 in Süddeutschland einen Gesamtschaden von etwa 3,6 Milliarden Euro (versicherte Schäden 2,8 Milliarden Euro; Tief Andreas; SwissRe, 2014; Kunz et al., 2018). Sehr schadenträchtig war auch eine Superzelle im Großraum München am 10. Juni 2019, die alleine zu Gesamtschäden von rund 1,0 Milliarde Euro führte (versicherte Schäden 0,75 Milliarden Euro; Munich Re, 2020; Wilhelm et al., 2021).

Starke konvektive Windböen, ebenfalls eine häufige Begleiterscheinung konvektiver Zellen, erreichen nicht selten Sturm- oder Orkanstärke (z. B. Mohr et al., 2017; Gatzen et al., 2020). Die Abwinde einer konvektiven Zelle werden in bodennahen Luftschichten horizontal umgelenkt und bilden eine Böenfront im Vorfeld einer Zelle. Seltener treten im Fall sehr starker Abwinde im Kern einer Zelle lokale, aber meist intensive Fallböen (*Downbursts*) auf. Das sogenannte Pfingstunwetter am 9. Juni 2014 beispielsweise verursachte in Nordrhein-Westfalen an einer ausgeprägten Böenfront in Form eines Bogenechos (*Bow Echos*) Orkanböen von lokal mehr als 140 $\mathrm{km\,h^{-1}}$ (Tief Ela; Barthlott et al., 2017; Mathias et al., 2017). Sechs Menschen starben infolge des Unwetters, und die Gesamtschäden betrugen in Deutschland knapp 0,9 Milliarden Euro (versicherte Schäden 0,65 Milliarden Euro; Munich Re, 2015).

Darüber hinaus werden in Deutschland jährlich etwa zwischen 20 und 60 Tornados beobachtet, die jedoch meist nur geringe Schäden verursachen (vgl. *European Severe Weather Database*, ESWD; Dotzek et al., 2009; Groenemeijer et al., 2017). Sie entstehen unterhalb einer Gewitterwolke im Bereich des Aufwinds durch die Generierung vertikaler Vorticity aus

horizontaler Vorticity, die aufgrund der vertikalen Scherung des Horizontalwinds in den bodennahen Schichten der Umgebung einer Zelle präsent ist. Vertikale Scherung bezeichnet dabei die Änderung der horizontalen Windgeschwindigkeit und -richtung mit der Höhe. Zu den verheerendsten Tornados in Deutschland im 20. und 21. Jahrhundert zählt ein Tornado, der am 10. Juli 1968 vom französischen Saartal ausgehend bis in die Region um Pforzheim zog. Er forderte in Pforzheim zwei Menschenleben und verursachte große Schäden an Gebäuden und Infrastrukturen (Nestle, 1969).

Kommt es zur Ausbildung hochreichender Feuchtkonvektion, hängt die weitere Entwicklung und damit der Lebenszyklus einer konvektiven Zelle von verschiedenen Faktoren ab, die in weiten Teilen gut, teils aber noch unzureichend verstanden sind. Die erste ausführliche Beschreibung des Lebenszyklus einer konvektiven Zelle findet sich bei Byers und Braham (1948), die diesen mit Hilfe einer Reihe von Beobachtungsdaten anhand verschiedener dynamischer und mikrophysikalischer Eigenschaften der Zellen in drei Stufen einteilen: 1) Cumulusstadium, 2) Reifestadium und 3) Dissipationsstadium. Die Autoren merkten bereits an, dass die weitere Entwicklung einer Zelle in Verbindung zu ihrem bisherigen Entwicklungsverlauf (der Zellhistorie) steht und ihr Fortbestehen durch atmosphärische Umgebungsbedingungen bestimmt werden könnte. Es folgten viele Studien, die sich intensiv mit der Beobachtung und der numerischen Modellierung konvektiver Zellen befassten (z. B. Orville und Sloan, 1970; Wilhelmson, 1974; Klemp und Wilhelmson, 1978b; Weisman und Klemp, 1982; Fovell und Tan, 1998).
Eine entscheidende Rolle für den Lebenszyklus konvektiver Zellen spielt deren Organisationsform (*Convective Mode*), welche ihrerseits eng mit den synoptisch- und mesoskaligen Umgebungsbedingungen in der Atmosphäre verbunden ist (z. B. Trapp, 2013). Basierend auf den beobachtbaren, strukturellen Charakteristika wird unterschieden zwischen isolierter Konvektion, bestehend aus einer einzelnen Konvektionszelle, multizellulärer Konvektion

und Mesoskaligen Konvektiven Systemen (MCS; horizontal über mehr als 100 km ausgedehnte, hochgradig multizelluläre Konvektion). Bei der isolierten Konvektion wird weiterhin zwischen eher kurzlebigen Einzelzellen und langlebigen Superzellen unterschieden. Die Grenzen zwischen den Organisationsformen sind dabei häufig fließend, da die Entwicklungen sehr dynamisch sind. Beispielsweise können zwei oder mehrere konvektive Zellen miteinander verschmelzen oder eine einzelne Zelle kann sich in zwei Zellen aufteilen. Letzteres wird vornehmlich bei Superzellen beobachtet. Während Einzelzellen meist keine allzu starke Entwicklung aufweisen und daher selten mit extremen Begleiterscheinungen einhergehen, können Superzellen aufgrund ihres hohen Grades an Organisation über mehrere Stunden bestehen, Zugbahnen mit einer Länge von mehreren hundert Kilometern aufweisen und daher große Schäden besonders durch großen Hagel, konvektive Windböen und Tornados verursachen – wie die oben erwähnten Superzellen in den Jahren 2013 und 2019. Auch Multizellen und MCS – wie das Pfingstunwetter 2014 – können über viele Stunden bis etwa einen Tag bestehen. Ihre Gefahr geht neben der Produktion von kleinem bis mittelgroßem Hagel und der Ausbildung intensiver Böenfronten besonders von lange anhaltendem, intensivem Starkregen aus.
Die Organisationsformen der konvektiven Zellen sind abhängig von den Umgebungsbedingungen, die anhand verschiedener meteorologischer Größen beschrieben werden können. Am wichtigsten ist dabei die vertikale Windscherung in der unteren und mittleren Troposphäre (z. B. Weisman und Klemp, 1982; Markowski und Richardson, 2010). Aber auch Maße zur Charakterisierung thermischer Instabilität der Atmosphärenschichtung sind von Relevanz. Mittlerweile gibt es eine Vielzahl von Umgebungsvariablen und spezieller konvektiver Indizes, welche die verschiedenen Voraussetzungen hochreichender Feuchtkonvektion zu quantifizieren versuchen und so eine Vorhersage konvektiver Zellen ermöglichen (*Ingredients-based Forecasting*; z. B. Huntrieser et al., 1997; Haklander und van Delden, 2003; Brooks, 2007; Kunz, 2007). Zudem

zeigen Studien, dass Prozesse auf der synoptischen Skala und Telekonnektionen die mesoskaligen Umgebungsbedingungen steuern können, die für die Gewitterentstehung relevant sind. Piper und Kunz (2017) zeigten beispielsweise, dass die Nordatlantische Oszillation über die Variation großskaliger Hebungsfelder in Verbindung mit synoptischen Wettersystemen einen Einfluss auf die konvektive Aktivität in Mitteleuropa hat. Atmosphärisches Blocking (Rex, 1950) über dem Baltikum begünstigt Wetterlagen mit hoher konvektiver Aktivität über West- und Mitteleuropa, die in der Regel mit einer eher geringen vertikalen Windscherung einhergehen (Mohr et al., 2019, 2020).

Trotz großer Fortschritte im Verständnis und in der numerischen Simulation konvektiver Zellen stellen diese aufgrund ihrer transienten und stochastischen Natur noch heute eine Herausforderung im Warnmanagement der Wetterdienste dar. Deren Warnungen basieren sowohl auf den Vorhersagen numerischer Wettervorhersagemodelle (NWV-Modelle) als auch auf aktuellen Beobachtungen des Wettergeschehens mittels verschiedener Datenquellen. Vor allem Daten von Fernerkundungsmethoden wie Messungen von Niederschlagsradaren, Satelliten und Blitzdetektionsnetzwerken sowie *in situ* Messungen von Wetterstationen, die nahezu in Echtzeit vorliegen, sind für die automatisierten Warnprozesse der Wetterdienste relevant.
Das Gefahrenpotential einer bevorstehenden Wetterlage kann auf der Basis der simulierten Umgebungsbedingungen und konvektiven Entwicklungen aus der NWV und daraus abgeleiteten Produkten häufig bereits ein bis drei Tage im Voraus erkannt werden, insbesondere wenn dabei auf probabilistische Vorhersagen zurückgegriffen wird (z. B. Gensini und Tippett, 2019). Die Wetterdienste können daher bereits frühzeitig auf eine mögliche Gefahrenlage aufmerksam machen. Großflächige Vorabinformationen zur Wetterlage werden ausgegeben, die aktualisiert werden, sobald sich verlässliche Aussagen zum genaueren Ablauf des Wettergeschehens treffen lassen. Handelt es sich um eine Wettersituation, die mit gefährlichen oder gar extremen Wetter- und

Folgeereignissen verbunden ist, werden dann nahezu in Echtzeit landkreis- oder postleitzahlenscharfe Warnungen in verschiedenen Warnstufen für die erwartete Dauer eines Ereignisses ausgegeben, welche die relevanten Wettererscheinungen und die damit verbundenen potentiellen Gefahren spezifizieren und Verhaltensvorschläge unterbreiten. Im Fall konvektiver Zellen wird zwischen den unterschiedlichen Begleiterscheinungen Starkregen, Hagel, Windböen und Tornados unterschieden, wobei abgestufte Hinweise auf die erwartete Niederschlagsmenge, Hagelkorngröße, Böenstärke bzw. das Auftreten eines Tornados gegeben werden. Solche Warnungen werden durch das *Nowcasting* der Wettersituation mit entsprechenden *Nowcasting*-Verfahren ermöglicht (Neologismus aus *now* [jetzt] und *forecasting* [Vorhersage]).

Nowcasting-Verfahren machen sich die aktuellen Beobachtungsdaten und NWV-Vorhersagen zunutze, um eine Abschätzung der Wetter- und Warnlage für die nächsten null bis zwei Stunden zu treffen. Dies ist alleine auf der Basis von operationellen Kürzestfristvorhersagen hochaufgelöster NWV-Modelle (null bis zwölf Stunden Vorhersagehorizont) bislang nicht realisierbar, da diese die neuesten Beobachtungen nicht berücksichtigen. Im Fall konvektiver Zellen kommt hinzu, dass selbst Vorhersagen hochaufgelöster Modelle, die Konvektion explizit simulieren, noch immer sehr große Unsicherheiten hinsichtlich des genauen Auftretens dieser Ereignisse aufweisen. *Nowcasting*-Verfahren hingegen können in teils hochkomplexen Wettersituationen über automatisierte Methoden, Strukturen und Abläufe Vorhersagen (*Nowcasts*) und Warnvorschläge mit einer höheren raum-zeitlichen Genauigkeit ausgeben. Qualifizierte Meteorolog*innen prüfen diese Warnvorschläge anhand einer kombinierten Interpretation der Beobachtungen, *Nowcasting*-Produkte und Vorhersagen aus der NWV, modifizieren diese gegebenenfalls und geben schließlich entsprechende Warnungen aus. Diese Warnungen sind nicht nur für Privatpersonen von Bedeutung, sondern auch für Einrichtungen des Katastrophenschutzes

sowie Entscheidungsträger in der Energie- und Landwirtschaft und im Veranstaltungsmanagement.

Der Deutsche Wetterdienst (DWD) realisiert seine automatische Unterstützung des Warnprozesses mit dem System AutoWARN (DWD, 2021e). Das *Nowcasting*-Verfahren NowCastMix findet dabei sowohl für das *Nowcasting* sommerlicher als auch winterlicher Warnereignisse Anwendung (James et al., 2018). NowCastMix kombiniert zur Erstellung eines Warnvorschlags NWV-Vorhersagen, Echtzeit-Meldungen von Wetterstationen, Blitzdaten, Wetterradarprodukte sowie Daten von Verfahren zur Erkennung und Verfolgung konvektiver Zellen (*Tracking*) mit einem Fuzzylogik-Ansatz. Im Rahmen des Projekts SINFONY (DWD, 2021c) gibt es zurzeit große Bemühungen, das bestehende *Nowcasting*-Verfahren und die Kürzestfristvorhersagen des hochaufgelösten NWV-Modells insbesondere mit dem Fokus auf sommerlicher Konvektion zu verbessern. Darüber hinaus soll ein integriertes Vorhersagesystem für den Zeitraum der Kürzestfrist entstehen, welches durch kombinierte Verfahren die *Nowcasts* und NWV-Kürzestfristvorhersagen homogenisiert und damit eine bruchfreie Vorhersage von Wetterereignissen mit Schadenpotential ermöglicht. Dadurch erhofft sich der DWD eine Verbesserung der Vorhersagequalität über den gesamten Kürzestfristzeitraum. Ähnliche Entwicklungen und Projekte gibt es derzeit auch in weiteren europäischen Ländern (Sun et al., 2014; Wapler et al., 2018, Schmid et al., 2019).
Auf der einen Seite soll in der NWV die Assimilation weiterer Beobachtungsdaten wie beispielsweise Radar-, Blitz- oder Satellitendaten in hoher zeitlicher Auflösung die Kürzestfristvorhersagen verbessern. Dazu wird ein sogenannter *Rapid Update Cycle* mit einer stündlichen Assimilation eingeführt. Wichtige Bestandteile für das *Nowcasting* konvektiver Zellen stellen auf der anderen Seite verbesserte und neue Verfahren zur Erkennung und Verfolgung konvektiver Zellen dar. Der Radarverbund des DWD liefert dazu

zwei- und dreidimensionale Messdaten (2D/3D) des Radarreflektivitätsfaktors über ganz Deutschland. Im Jahr 2000 führte der DWD den Zellverfolgungsalgorithmus KONRAD (Konvektionsentwicklung in Radarprodukten) ein, der in 2D-Radarprodukten konvektive Zellen erkennt und zeitlich verfolgt. Als Nachfolge von KONRAD wird momentan die Neuentwicklung des Systems KONRAD3D präoperationell getestet (Werner, 2020). Dieses Verfahren berücksichtigt 3D-Radarmessungen und kann auf neu entwickelte Techniken zur Qualitätssicherung von Radardaten und zur quantitativen Niederschlagsabschätzung zurückgreifen sowie eine Methodik zur Hydrometeorklassifikation ausnutzen.
Zur Weiterentwicklung von KONRAD3D und einer potentiellen Verbesserung der *Nowcasts* und der Warnvorschläge sind statistische Analysen von Daten des bestehenden KONRAD-Verfahrens und anderer Zellverfolgungsalgorithmen nützlich (z. B. Wapler, 2017; Zöbisch et al., 2020). Die hier gewonnenen Informationen über verschiedene Eigenschaften der konvektiven Zellen werden dazu verwendet, ein statistisches Modell für den Lebenszyklus der Zellen zu entwickeln (z. B. Feger et al., 2019; Zöbisch, 2020; Wapler, 2021). Ein solches Modell ermöglicht eine Abschätzung der Verlagerung, Intensitätsentwicklung und Lebensdauer konvektiver Zellen sowie möglicher Begleiterscheinungen unabhängig von der Organisationsform. Bei dieser Abschätzung besteht jedoch aufgrund der großen Variabilität der Entwicklungen konvektiver Zellen in den unterschiedlichen Organisationsformen weiterhin ein großes Verbesserungspotential. Von besonderem Interesse ist daher im Rahmen der Entwicklung integrierter Vorhersageverfahren auch die Möglichkeit, Informationen aus den NWV-Vorhersagen und/oder weitere Beobachtungsdaten direkt in einem Zellverfolgungsalgorithmus wie KONRAD zu nutzen (Multi-Sensor-/Multi-Daten-Ansatz; z. B. Wapler et al., 2015; Josipovic, 2020). Dadurch erhoffen sich die Wetterdienste eine verbesserte und probabilistische Abschätzung des Lebenszyklus konvektiver Zellen (Wapler et al., 2018; Schmid et al., 2019).

Die Zielsetzung der vorliegenden Arbeit besteht darin, aus der Kombination von Daten des Zellverfolgungsalgorithmus KONRAD für konvektive Zellen mit hochaufgelösten Modelldaten des DWD eine kombinierte statistische Analyse des Lebenszyklus konvektiver Zellen im Sinne des Multi-Daten-Ansatzes vorzunehmen. Anschließend geht diese Arbeit einen Schritt weiter und entwickelt und evaluiert Verfahren, die auf der Basis dieser Statistik eine verbesserte Abschätzung bestimmter Eigenschaften des Lebenszyklus ermöglichen. Diese Verfahren haben das Potential, das *Nowcasting* in einem integrierten Vorhersagesystem durch das Einbeziehen von NWV-Vorhersagen in einen Zellverfolgungsalgorithmus zu verbessern.
Dazu müssen zunächst geeignete Lebenszyklen aus den Rohdaten des Zellverfolgungsalgorithmus generiert werden. Außerdem ist es notwendig, verschiedene potentiell relevante Umgebungsvariablen und konvektive Indizes aus den Modelldaten zu bestimmen. Darauf aufbauend erfolgt die Erstellung eines kombinierten Datensatzes, der aus Lebenszyklen konvektiver Zellen und einer großen Anzahl von Umgebungsvariablen für sechs Sommerhalbjahre (2011 – 2016) über Deutschland und angrenzenden Regionen besteht. Die Lebenszyklen können nachfolgend alleine oder in Kombination mit den vorherrschenden Umgebungsbedingungen analysiert werden. Diese Analysen identifizieren statistische Zusammenhänge und setzen sie in den physikalisch-meteorologischen Kontext. Der kombinierte Datensatz bildet zudem die Grundlage für die Anwendung mehrerer Verfahren der Statistik und des maschinellen Lernens zur Abschätzung bestimmter Eigenschaften der Lebenszyklen. Im Detail wird untersucht, wie gut ein multivariates, logistisches Regressionsmodell, ein nicht-linearer Polynomansatz und ein *Random Forest* die weitere Lebensdauer und die maximale (horizontale) Zellfläche von konvektiven Zellen auf der Basis der Umgebungsvariablen und der Zellhistorie vorhersagen können. Diese Verfahren liefern unter Anwendung eines Ensembleansatzes eine probabilistische Abschätzung dieser beiden Zellattribute im Sinne einer Klassifikation (z. B. kurze/lange Lebensdauer) oder einer Regression (z. B. Lebensdauer in Minuten).

Dementsprechend orientieren sich die Untersuchungen der vorliegenden Arbeit an den folgenden fünf zentralen Fragestellungen:

(1) Wie gut lassen sich Lebenszyklen konvektiver Zellen aus Daten eines Zellverfolgungsalgorithmus extrahieren und wo liegen die Grenzen?

(2) Welche statistischen Eigenschaften weisen diese Lebenszyklen konvektiver Zellen in Deutschland auf?

(3) Unter welchen Umgebungsbedingungen treten konvektive Zellen in Deutschland auf und welche statistischen Zusammenhänge lassen sich zwischen ihnen erkennen?

(4) Ist es möglich, den Lebenszyklus konvektiver Zellen im Sinne einer probabilistischen Vorhersage auf der Grundlage der statistischen Analysen (Punkte 2 und 3) besser abzuschätzen?

(5) Welches Verfahren und welche Parameter eignen sich am besten zur potentiellen Ergänzung bestehender *Nowcasting*-Verfahren?

Die vorliegende Arbeit wird neben der Lebenszyklusanalyse konvektiver Zellen das Potential für eine Verbesserung der Vorhersage des Lebenszyklus durch das Einbeziehen von NWV-Vorhersagen in einen Zellverfolgungsalgorithmus aufzeigen. Der Vergleich mehrerer Verfahren aus der Statistik und des maschinellen Lernens verdeutlicht, dass die Wahl eines geeigneten Verfahrens von der genauen Fragestellung bzw. Anforderung abhängt.

Der theoretische, physikalisch-meteorologische Hintergrund in Kapitel 2 leitet in die Thematik hochreichender Feuchtkonvektion und der Lebenszyklen unterschiedlicher konvektiver Systeme ein. In Kapitel 3 schließt sich eine mathematische Darstellung der verwendeten Verfahren aus der Statistik und des maschinellen Lernens an. Hinzu kommen Beschreibungen verschiedener

Evaluationsmaße. Kapitel 4 stellt im Anschluss die Datengrundlage vor (Daten aus dem Zellverfolgungsalgorithmus, Daten aus dem NWV-Modell). Außerdem werden die Methoden der Datenaufbereitung erläutert, die zur Erstellung eines kombinierten Datensatzes führen. Kapitel 5 präsentiert verschiedene Analysen dieses Datensatzes und diskutiert die Zusammenhänge zwischen den Lebenszyklen konvektiver Zellen und den vorherrschenden Umgebungsbedingungen. Die Entwicklung und Evaluation der verschiedenen Verfahren zur Abschätzung der Lebensdauer und der maximalen Zellfläche konvektiver Zellen ist Inhalt von Kapitel 6. Abschließend folgt in Kapitel 7 eine Zusammenfassung und Diskussion der wichtigsten Ergebnisse.

2 Theoretischer Hintergrund und thematische Einordnung

Die Atmosphäre der Erde ist im physikalischen Sinne ein Fluid, welches sowohl dynamischen Einflüssen aufgrund des Wirkens unterschiedlicher Kräfte als auch thermodynamischen Modifikationen aufgrund von Energieumwandlungen ausgesetzt ist (Vallis, 2017). Die kausalen Wirkungsketten atmosphärischer Prozesse weisen zudem eine Interaktion von nicht-linearen dynamischen und thermodynamischen Vorgängen auf, sodass diese einer gemeinsamen Beschreibung bedürfen, wenn man alle Komponenten des Systems Atmosphäre berücksichtigen möchte. Um die Wirkung der komplexen atmosphärischen Prozesse zu verstehen, ist eine grundlegende Betrachtungsweise einzelner physikalischer Prinzipien unerlässlich. Die nachfolgenden theoretischen Herleitungen und Erläuterungen fokussieren sich dabei auf thermodynamische Grundlagen, die für die Entstehung von hochreichender Feuchtkonvektion relevant sind (Kapitel 2.1). Welche verschiedenen Gewittersysteme beobachtet werden und worin sich diese hinsichtlich verschiedener Charakteristika unterscheiden, legt Kapitel 2.2 dar. Dabei werden einige atmosphärische Variablen vorgestellt, mit denen sich die atmosphärischen Umgebungsbedingungen hinsichtlich der bevorzugten Entstehung unterschiedlicher konvektiver Systeme beschreiben lassen. Kapitel 2.3 stellt anschließend weitere relevante Umgebungsvariablen und spezielle Kenngrößen vor, die verschiedene dynamische und thermodynamische Aspekte quantifizieren, während Kapitel 2.4 abschließend eine Brücke zwischen den meteorologischen Grundlagen und den Charakteristika der Lebenszyklen konvektiver Systeme schlägt.

2.1 Entstehungsmechanismen hochreichender Konvektion

2.1.1 Adiabatische Zustandsänderungen in der Atmosphäre

Ein thermodynamisches System ist ein räumlich abgrenzbares System mit physikalischen Eigenschaften, die durch die Gesetze der Thermodynamik beschrieben werden können. Ein solches System ist durch seine Zustandsgrößen charakterisiert, welche sowohl extensiv als auch intensiv sein können. Die Werte von extensiven Zustandsvariablen sind proportional zur Größe des Systems, welche durch die Skalierungsmaße Teilchenanzahl N oder Masse m beschrieben wird. Die Werte von intensiven Zustandsvariablen hingegen sind unabhängig von N bzw. m. Zu den extensiven Zustandsgrößen zählen beispielsweise die Entropie S, das Volumen V oder die innere Energie U, zu den intensiven Zustandsgrößen der Druck p und die Temperatur T des Systems. Im thermodynamischen Gleichgewicht, einem Zustand ohne makroskopische Energie- und Massenflüsse, kann U als eine Funktion der weiteren extensiven Eigenschaften ausgedrückt werden (z. B. Vallis, 2017). Dividiert man diese Eigenschaften, wie in der Fluiddynamik üblich, durch die Masse des Fluids, so lässt sich ein funktionaler Zusammenhang wie folgt formulieren:

$$u \equiv u(\alpha, s, \mathbf{b}) \ . \tag{2.1}$$

Darin steht $\alpha = V\,m^{-1} = \rho^{-1}$ für das spezifische Volumen, ρ kennzeichnet die Dichte und s die spezifische Entropie des Fluids. Die Variable $\mathbf{b}$ parametrisiert die spezifischen Anteile der Bestandteile des Fluids. Alle Variablen stellen nun intensive Eigenschaften des Systems dar. Im Fall einer feuchten, ungesättigten Atmosphäre mit den Komponenten trockene Luft und Wasserdampf existieren laut der Gibbs'schen Phasenregel für ein zweikomponentiges ($K = 2$), einphasiges ($P = 1$) System $F = K - P + 2 = 3$ Freiheitsgrade, welche die Anzahl der unabhängigen intensiven Eigenschaften beschreiben.

Die Parametrisierung **b** kann hier als Anteil der trockenen Luft oder des Wasserdampfs an der Gesamtzahl von Molekülen verstanden werden, der jeweils andere Anteil ergibt sich zwangsläufig aus dem ersten. Im Fall einer trockenen Atmosphäre ist wegen $K = 1$ die Anzahl der Freiheitsgrade $F = 2$.
Änderungen des Zustands eines thermodynamischen Systems und damit auch der Atmosphäre folgen dem physikalischen Prinzip der Energieerhaltung, welches im Ersten Hauptsatz der Thermodynamik Ausdruck findet. Für ein geschlossenes System, dessen Zusammensetzung sich während der thermodynamischen Prozessführung nicht ändert, lautet dieser:

$$\mathrm{d}u = \delta q + \delta w \; . \tag{2.2}$$

Das Hinzu- bzw. Abführen von Wärme δq sowie das Verrichten von Arbeit am System δw können demnach zu einer Änderung der inneren Energie du führen. Das vorangestellte d kennzeichnet ein vollständiges Differential, während δ ein wegabhängiges Differential darstellt. Die verrichtete Arbeit stellt die Volumenarbeit am System dar, sodass $\delta w = -p\mathrm{d}\alpha$ gilt. Mit der Definition der spezifischen Enthalpie des Systems $h = u + p\alpha$ lässt sich Gleichung (2.2) unter Anwendung der Produktregel daher über

$$\mathrm{d}h = \delta q + \alpha \mathrm{d}p \tag{2.3}$$

darstellen.
Adiabatische Zustandsänderungen in einem geschlossenen System beschreiben Änderungen ohne Zu- oder Abfuhr von Wärme, d. h. $\delta q = 0$. Sie stellen eine oftmals verwendete Näherung für das meteorologische Konzept eines gehobenen Luftpakets dar, welches zur Beschreibung des Mechanismus von Konvektion Anwendung findet (s. u.; z. B. Bjerknes, 1938; Markowski und Richardson, 2010). In einer trockenen Atmosphäre gilt unter der berechtigten Annahme, dass sich die Luft in der Atmosphäre wie ein

ideales Gas verhält, folgende Zustandsgleichung:

$$p\alpha = p_d\alpha = R_d T \ . \tag{2.4}$$

Der Index d (*dry*) charakterisiert fortan Größen, die sich auf trockene Luft beziehen, wobei im Fall einer trockenen Atmosphäre der Partialdruck trockener Luft gleich dem Atmosphärendruck ist: $p_d = p$. In Gleichung (2.4) kennzeichnet $R_d = 287{,}05\,\mathrm{J\,kg^{-1}\,K^{-1}}$ die Gaskonstante für trockene Luft. Hier wird deutlich, dass ein Zustand eindeutig durch zwei der drei Zustandsgrößen Druck, Temperatur und spezifisches Volumen gegeben ist. Nach Joules zweitem Gesetz (z. B. Tsonis, 2007) ist die innere Energie allein von der Temperatur abhängig, d. h. $u = u(T)$ und somit gilt auch für die Enthalpie unter Verwendung der Zustandsgleichung (2.4): $h = u(T) + R_d T = h(T)$. Mit der spezifischen Wärmekapazität bei konstantem Druck

$$c_p = \left(\frac{\partial h}{\partial T}\right)_p = \frac{\mathrm{d}h}{\mathrm{d}T} \ , \tag{2.5}$$

wobei $c_p = 1\,004{,}5\,\mathrm{J\,kg^{-1}\,K^{-1}}$ gilt, folgt daher aus Gleichung (2.3)

$$c_p\mathrm{d}T = \delta q + \alpha\mathrm{d}p \ . \tag{2.6}$$

Ein trockenadiabatischer Prozess wird demnach durch $c_p\mathrm{d}T = \alpha\mathrm{d}p$ oder mit Hilfe der Zustandsgleichung (2.4) und der Abkürzung $\kappa \equiv R_d c_p^{-1}$ umformuliert über

$$\mathrm{d}\ln(T) = \kappa\mathrm{d}\ln(p) \tag{2.7}$$

beschrieben. Die Einführung des Adiabatenkoeffizienten $\gamma = c_p c_V^{-1}$ mit der spezifischen Wärmekapazität bei konstantem Volumen

$$c_V = \left(\frac{\partial u}{\partial T}\right)_V = \frac{\mathrm{d}u}{\mathrm{d}T} = c_p - R_d \tag{2.8}$$

führt auf $\kappa = (\gamma - 1)\ \gamma^{-1}$. Die unbestimmte Integration von Gleichung (2.7) liefert schließlich eine der bekannten Poisson-Gleichungen für adiabatische Prozesse:

$$T^{\gamma} p^{1-\gamma} = const. \tag{2.9}$$

Die Definition der potentiellen Temperatur θ erhält man als eine Realisierung dieser Poisson-Gleichung durch die analytische Integration von Gleichung (2.7) von einem Referenzniveau p_0 bis zu einem unbestimmten Druckniveau p:

$$\theta(T,p) \equiv T(p_0) = T \left(\frac{p_0}{p} \right)^{\kappa} . \tag{2.10}$$

Üblicherweise wird $p_0 = 1000\,\mathrm{hPa}$ verwendet. Die potentielle Temperatur $\theta(T,p)$ kann demnach als die Temperatur verstanden werden, die ein trockenes Luftpaket auf einem Druckniveau p mit der Temperatur T annehmen würde, wenn es trockenadiabatisch auf das Druckniveau p_0 absinken würde. Daher ist $\theta(T,p)$ invariant unter trockenadiabatischen Zustandsänderungen.

Im Fall feuchter, ungesättigter Luft und unter der Annahme, dass sich der in der Atmosphäre vorhandene Wasserdampf wie ein ideales Gas verhält, lässt sich die Zustandsgleichung (2.4) gemäß des Dalton'schen Gesetzes leicht erweitern, da sich der Druck der Atmosphäre p als Summe der Partialdrücke der beiden Gase ergibt (z. B. Markowski und Richardson, 2010):

$$p = p_d + p_v = p_d + e = (R_d \rho_d + R_v \rho_v) T\ . \tag{2.11}$$

Darin kennzeichnet p_v den Wasserdampfpartialdruck, der in der Literatur meist mit e bezeichnet wird. Außerdem stellt $R_v = 461{,}51\,\mathrm{J\,kg^{-1}\,K^{-1}}$ die Gaskonstante von Wasserdampf und ρ_v dessen Dichte dar (Index v: *vapor*).

Analog zur Herleitung von Gleichung (2.7) lässt sich für einen trockenadiabatischen Prozess feuchter Luft ohne Phasenumwandlungen

$$\mathrm{dln}(T) = \kappa_u(r_{v,0})\mathrm{dln}(p) \tag{2.12}$$

mit $\kappa_u(r_{v,0}) = (R_d + r_{v,0}R_v)(c_{p,d} + r_{v,0}c_{p,v})^{-1}$ und dem konstanten Mischungsverhältnis von Wasserdampf zu trockener Luft $r_{v,0} = \rho_v\rho_d^{-1}$ schreiben (Manzato und Morgan, 2003). Die spezifische Wärmekapazität von Wasserdampf beträgt $c_{p,v} = 1\,845{,}6\,\mathrm{J\,kg^{-1}\,K^{-1}}$. Die entsprechende Invariante ist

$$\theta_u(T, p, r_{v,0}) = T\left(\frac{p_0}{p}\right)^{\kappa_u(r_{v,0})} . \tag{2.13}$$

Häufig wird zur Vermeidung der Abhängigkeit von $r_{v,0}$ auf ein alternatives Temperaturmaß zurückgegriffen, das es im Fall einer feuchten, ungesättigten Atmosphäre ermöglicht, formal mit einer trockenen Atmosphäre zu rechnen. Die sogenannte virtuelle Temperatur

$$T_V = T\,\frac{r_v + \frac{R_d}{R_v}}{\frac{R_d}{R_v}(1+r_v)} \approx T\left(1+0{,}608\frac{r_v}{1+r_v}\right) \overset{r_v \ll 1}{\approx} T\,(1+0{,}608 r_v) \tag{2.14}$$

stellt die Temperatur dar, die ein trockenes Luftpaket haben müsste, um dieselbe Dichte wie feuchte Luft bei gleichem Druck zu haben. Die entsprechende Zustandsgleichung für ein ideales Gas lautet

$$p\alpha = R_d T_V \tag{2.15}$$

und die virtuelle potentielle Temperatur θ_V ergibt sich analog zu Gleichung (2.10) über

$$\theta_V(T_V, p) = T_V\left(\frac{p_0}{p}\right)^{\kappa} . \tag{2.16}$$

Im Fall feuchter, gesättigter Luft, in der Phasenumwandlungen auftreten, gestaltet sich die Bestimmung einer Invarianten deutlich schwieriger. Es sei zunächst ein Prozess betrachtet, bei dem die frei werdende latente Wärme aus der Kondensation des Wasserdampfs in einem Luftpaket komplett zu dessen Heizung verwendet wird (Holton, 2004; Vallis, 2017). Da kein Wärmeaustausch mit der Umgebung stattfindet, gilt weiterhin $\delta q = 0$. In Anlehnung an Simpson (1978), Holton (2004), Markowski und Richardson (2010) und Vallis (2017) kann Gleichung (2.7) jedoch durch eine interne Heizrate $\delta q_{int} = -T\ \mathrm{d}(l_v(T) r_{v,s}\ T^{-1})$ modifiziert werden, um die Kondensationsprozesse im Luftpaket näherungsweise abzubilden:

$$\mathrm{dln}(T) = \kappa\ \mathrm{dln}(p_d) + \frac{\kappa}{R_d T} \delta q_{int}\ . \tag{2.17}$$

Dabei steht $l_v(T)$ für die spezifische Verdampfungswärme von Wasser, die für typische Temperaturen in der Atmosphäre zwischen $\vartheta = -50$ und $30\,°\mathrm{C}$ um weniger als $10\,\%$ variiert. Empirisch bestimmt wurde etwa $l_v(\vartheta = 0\,°\mathrm{C}) \approx 2{,}501 \cdot 10^6\,\mathrm{J\,kg^{-1}}$. In δq_{int} kennzeichnet $r_{v,s}(p,T)$ das Sättigungsmischungsverhältnis von Wasserdampf beim entsprechenden Sättigungsdampfdruck $e_s(T)$, welcher den Wasserdampfpartialdruck bei Sättigung der Luft angibt und über die Clausius-Clapeyron-Gleichung (z. B. Seinfeld und Pandis, 2006) beschrieben werden kann:

$$\frac{\mathrm{d}e_s(T)}{\mathrm{d}T} = \frac{l_v(T) e_s(T)}{R_v T^2}\ . \tag{2.18}$$

Die Beziehung zwischen dem Mischungsverhältnis und dem entsprechenden Wasserdampfpartialdruck e ergibt sich durch die Kombination der Zustandsgleichungen für trockene (2.4) und feuchte, ungesättigte Luft (2.11):

$$r_v = \frac{\rho_v}{\rho_d} = \frac{R_d e}{R_v p_d} \quad \Longrightarrow \quad r_{v,s}(T,p) = \frac{R_d e_s(T)}{R_v(p - e_s(T))}\ . \tag{2.19}$$

Gleichung (2.17) ist identisch mit

$$\mathrm{d}\ln(\theta_d) = -\frac{1}{c_{p,d}}\mathrm{d}\left(\frac{l_v(T)r_{v,s}}{T}\right) , \tag{2.20}$$

wobei θ_d über Gleichung (2.10) mit der Ersetzung $p \to p_d$ gegeben ist. Durch Integration vom Ausgangszustand (θ_d,$r_{v,s}$,T) bis zu einem Zustand, in dem die Luft nahezu keinen Wasserdampf mehr enthält ($r_{v,s} \approx 0$), ergibt sich schließlich die Invariante

$$\begin{aligned}\theta_e(T,p,e) &= \theta_d \exp\left(\frac{l_v r_{v,s}(T,p)}{c_{p,d}T}\right) \\ &= T\left(\frac{p_0}{p-e}\right)^{\kappa} \exp\left(\frac{l_v r_{v,s}(T,p)\kappa}{R_d T}\right) , \end{aligned} \tag{2.21}$$

welche als äquivalentpotentielle Temperatur bezeichnet wird. Da $p \gg e$ ist, wird häufig $p - e \approx p$ gesetzt, sodass sich Gleichung (2.21) zu $\theta_e = \theta_e(T,p)$ vereinfacht.

Simpson (1978) und Markowski und Richardson (2010) merken an, dass für einen reversiblen Prozess, in dem das kondensierte Wasser weiterhin im Luftpaket enthalten ist, in Gleichung (2.17) κ durch $\kappa_r(r_{w,0}) = R_d(c_{p,d} + r_{w,0}c_{p,l})^{-1}$ ersetzt werden sollte, um den Effekt des Wasserdampfs und des flüssigen Wassers auf die spezifische Wärmekapazität des Luftpakets zu berücksichtigen. Dabei ist $r_{w,0}$ das konstante Mischungsverhältnis von Wasserdampf und flüssigem Wasser zu trockener Luft und $c_{p,l} = 4\,218\,\mathrm{J\,kg^{-1}\,K^{-1}}$ die spezifische Wärmekapazität von flüssigem Wasser. Damit folgt für die äquivalentpotentielle Temperatur unter der Annahme eines reversiblen Prozesses durch eine analoge Rechnung

$$\theta_e^{(r)}(T,p,e,r_{w,0}) = T\left(\frac{p_0}{p-e}\right)^{\kappa_r(r_{w,0})} \exp\left(\frac{l_v r_{v,s}(T,p)\kappa_r(r_{w,0})}{R_d T}\right) . \tag{2.22}$$

$\theta_e^{(r)}(T,p,e,r_{w,0})$ ist folglich näherungsweise invariant unter feuchtadiabatischen Zustandsänderungen ohne Massenänderungen, d. h. ohne ausfallenden Niederschlag.
Für einen irreversiblen Prozess in feuchter, gesättigter Luft, bei dem kondensiertes flüssiges Wasser instantan vollständig aus einem Luftpaket entfernt wird, ist die adiabatische Annahme $\delta q = 0$ nicht mehr gerechtfertigt. Dieser Prozess, welcher allgemein als pseudoadiabatisch bezeichnet wird, lässt sich in zwei Stufen vorstellen: Zunächst erfolgt eine irreversible, feuchtadiabatische Expansion, die zur Kondensation führt. Anschließend wird dem Luftpaket das kondensierte Wasser unter Erhalt von Temperatur und Druck entzogen, was die Entropie reduziert. Simpson (1978) erläutert, dass in Gleichung (2.17) in einem solchen Prozess κ durch $\kappa_i(r_{v,s}) = R_d(c_{p,d} + r_{v,s}c_{p,l})^{-1}$ ersetzt werden muss, sodass gilt:

$$\mathrm{dln}(T) = \kappa_i(r_{v,s})\mathrm{dln}(p_d) + \frac{\kappa_i(r_{v,s})}{R_d T}\delta q_{int} \tag{2.23}$$

$$\Longleftrightarrow \quad \mathrm{dln}(\theta_d) = -\frac{\kappa_i(r_{v,s})}{R_d}\left[\mathrm{d}\left(\frac{l_v(T)r_{v,s}}{T}\right) + r_{v,s}c_{p,l}\mathrm{dln}T\right] . \tag{2.24}$$

Da aber $r_{v,s} = r_{v,s}(T,p)$ ist, lässt sich der letzte Term in Gleichung (2.24) nur aufwändig numerisch integrieren. Eine sehr gute Näherungsformel für die Invariante entwickelte Bolton (1980):

$$\begin{aligned}\theta_{ps} &= \theta_e^{(i)}(T,p,e)\\ &\approx T\left(\frac{p_0}{p}\right)^{a_1[1-a_2 r_v(p,e)]^{-1}}\\ &\quad\cdot\exp\left[\left(\frac{a_3}{T_{\mathrm{HKN}}(T,e)} - a_4\right) r_v(p,e)[1 + a_5 r_v(p,e)]\right] .\end{aligned} \tag{2.25}$$

Darin erscheinen mehrere numerische Konstanten a_k und ein empirisch bestimmter Zusammenhang zwischen der ungefähren Temperatur in der Höhe des Hebungskondensationsniveaus $T_{\mathrm{HKN}}(T,e)$ und der Temperatur T sowie

dem Wasserdampfpartialdruck e (s. Kapitel 2.1.2). $\theta_e^{(i)}$ wird auch als pseudopotentielle Temperatur θ_{ps} bezeichnet und stellt das oftmals verwendete Temperaturmaß zur Beschreibung feuchtadiabatischer Prozesse dar, so auch in der vorliegenden Arbeit.

2.1.2 Vertikalbewegungen in der Atmosphäre

Vertikalbeschleunigung und Auftrieb

Die dynamischen und thermodynamischen Prozesse der Atmosphäre können durch ein gekoppeltes, nicht-lineares Gleichungssystem beschrieben werden, die Ausdruck der Impuls-, Massen- und Energieerhaltung in einem rotierenden, geschlossenen System sind. Die entsprechenden prognostischen Variablen hängen dabei von dem betrachteten Ort $\mathbf{x}$ und dem Zeitpunkt t ab, sodass die totale Zeitableitung durch die materielle Ableitung

$$\frac{\mathrm{d}}{\mathrm{d}t} \quad \longrightarrow \quad \frac{\mathrm{D}}{\mathrm{D}t} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \tag{2.26}$$

mit dem Standardskalarprodukt $(\cdot)$ und dem Nabla-Operator (∇) ausgedrückt werden kann:

$$\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = \underbrace{-\nabla p}_{(1)} \underbrace{+\rho \mathbf{g}'}_{(2)} \underbrace{-2\rho \boldsymbol{\Omega} \times \mathbf{v}}_{(3)} \underbrace{-\rho \nabla \cdot \mathscr{T}}_{(4)} \tag{2.27}$$

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} = -\rho \nabla \cdot \mathbf{v} \tag{2.28}$$

$$\frac{\mathrm{D}\theta}{\mathrm{D}t} = \frac{\rho \theta \kappa}{p} \frac{\mathrm{D}q}{\mathrm{D}t} \ . \tag{2.29}$$

Darin kennzeichnet das Kreuz $(\times)$ das Kreuzprodukt zweier Vektoren. Diese fünf Gleichungen stellen zusammen mit der Zustandsgleichung für ein ideales Gas (2.4) ein Gleichungssystem für eine trockene Atmosphäre mit folgenden

prognostischen atmosphärischen Variablen dar: Windvektor $\mathbf{v} = \mathbf{u} + w\mathbf{e}_z$, bestehend aus dem Horizontalwind $\mathbf{u}$ und der vertikalen Komponente w, Luftdruck p und potentielle Temperatur θ. Die erste Gleichung ist die Impulsgleichung in der Form einer Bilanzgleichung, welche sich durch das Wirken (1) der Druckgradientkraft, (2) der scheinbaren Erdanziehungskraft[1], (3) der Corioliskraft und (4) Reibungskräften ergibt. Die zweite Gleichung ist die (Massen-)Kontinuitätsgleichung und die dritte die thermodynamische Energiegleichung, welche sich direkt aus dem Ersten Hauptsatz der Thermodynamik ergibt, wie beispielsweise aus Gleichung (2.20) mit einer allgemeinen Heizrate δq statt $\delta q = -T\mathrm{d}(l_v r_{v,s} T^{-1})$. Für eine Atmosphäre mit Wasser in allen Aggregatzuständen erfolgt eine Erweiterung dieses Gleichungssatzes um weitere Terme in den obigen Gleichungen sowie um Tendenzgleichungen für den Wasserdampfgehalt (spezifische Feuchte), den spezifischen Flüssigwassergehalt (Wolken- und Regentropfen) und den spezifischen Gehalt gefrorenen Wassers (Eis, Graupel, Schnee), welche die entsprechenden Quellen und Senken sowie Diffusionsprozesse berücksichtigen. Im obigen Gleichungssystem steht $\mathbf{\Omega}$ für den konstanten Vektor der Winkelgeschwindigkeit der Erdrotation und $\mathscr{T}$ für den viskosen Spannungstensor (z. B. Vallis, 2017). Zur Beschreibung von Vertikalbewegungen, wie beispielsweise denen von aufsteigenden Luftpaketen, dient die vertikale Komponente der Impulsgleichung (2.27) unter Vernachlässigung von Effekten durch die Corioliskraft und Reibungskräfte, mit der Näherung $\mathbf{g}' \approx \mathbf{g} \approx -g\ \mathbf{e}_z$ mit $g \approx 9{,}81\,\mathrm{m\,s^{-1}}$ für mittlere Breiten (Breitengrad $\phi = 45°$):

$$\rho \frac{\mathrm{D}w}{\mathrm{D}t} \approx -\frac{\partial p}{\partial z} - \rho g\ . \tag{2.30}$$

[1] Die scheinbare Erdbeschleunigung $\mathbf{g}'$ berücksichtigt sowohl die Geoidform der Erde als auch die aufgrund der Erdrotation wirkende Zentrifugalkraft.

Eine sehr häufig verwendete Beschreibung der Atmosphäre zerlegt die Beiträge der Zustandsvariablen in einen horizontal homogenen Grundzustand $(\overline{p}(z), \overline{\rho}(z))$ und Fluktuationen $(p'(\mathbf{x},t), \rho'(\mathbf{x},t))$ um diesen herum, auch Störungen genannt. Für den Grundzustand wird zudem Hydrostasie angenommen, d. h.

$$\frac{\partial \overline{p}}{\partial z} = -\overline{\rho} g \; , \tag{2.31}$$

sodass aus der Impulsgleichung (2.30) folgt:

$$\frac{\mathrm{D}w}{\mathrm{D}t} \approx -\frac{1}{\rho}\frac{\partial p'}{\partial z} + B \; . \tag{2.32}$$

Darin steht $B = -g\rho'\rho^{-1}$ für den thermischen Auftrieb (*Buoyancy*). Im Allgemeinen ist $\rho' \ll \overline{\rho}$, sodass in feuchter, ungesättigter Luft gilt:

$$B_u \approx -\frac{\rho'}{\overline{\rho}} g \overset{(2.15)}{\approx} \left(\frac{T_V'}{\overline{T}_V} - \frac{p'}{\overline{p}} \right) g \; . \tag{2.33}$$

Zur Berücksichtigung von Hydrometeoren in der flüssigen und festen Phase im Fall gesättigter Luft ist der thermische Auftrieb nach Markowski und Richardson (2010) auf

$$B_s \approx -\frac{\rho'}{\overline{\rho}} g \overset{(2.15)}{\approx} \left(\frac{T_V'}{\overline{T}_V} - \frac{p'}{\overline{p}} - r_h \right) g \tag{2.34}$$

zu erweitern, wobei r_h für das Mischungsverhältnis der Hydrometeore im Gesamten steht. Diese wirken aufgrund ihres Gewichts dem Auftrieb entgegen. Markowski und Richardson (2010) und Trapp (2013) zeigen darüber hinaus anhand der Impulsgleichung (2.27) unter einigen geeigneten Annahmen, dass der vertikale Gradient der Druckstörungen in einen durch den Auftrieb bedingten (p_b') und einen dynamisch hervorgerufenen Anteil (p_{dyn}') zerlegt werden kann:

$$\frac{\mathrm{D}w}{\mathrm{D}t} \approx -\frac{1}{\overline{\rho}}\frac{\partial p_{dyn}'}{\partial z} + \left(B - \frac{1}{\overline{\rho}}\frac{\partial p_b'}{\partial z} \right) \; . \tag{2.35}$$

Dabei hängt p'_b direkt mit dem vertikalen Gradienten des Auftriebs B zusammen: Oberhalb (unterhalb) eines aufsteigenden Luftpakets führt ein negativer (positiver) Gradient zu einer positiven (negativen) Druckanomalie, die als Verdrängen (Nachströmen) der Luft der Umgebung interpretiert werden kann. Dadurch induziert das Luftpaket jedoch eine zusätzliche, abwärts gerichtete Druckgradientkraft, die dem thermischen Auftrieb B entgegenwirkt. Diese ist umso stärker, je horizontal ausgedehnter das Luftpaket ist. Das Einmischen (*Entrainment*) von (trockenerer, kälterer) Luft aus der Umgebung in das Luftpaket trägt ebenfalls zur Verringerung des thermischen Auftriebs B bei (z. B. Lohmann et al., 2016). Der Beitrag von p'_{dyn} äußert sich in verschiedenartigen Deformationen des Luftpakets durch das Strömungsfeld $\mathbf{v}$.

Theorie eines gehobenen Luftpakets

Anhand dieser Beschreibung wird bereits deutlich, dass für ein aufsteigendes Luftpaket in der Atmosphäre die oben getroffene Annahme eines adiabatischen Prozesses, in dem weder Energie- noch Massenaustausch mit der Umgebung stattfindet, auch ohne Phasenumwandlungsprozesse von Wasser eine starke Vereinfachung der Realität darstellt. In der klassischen, konzeptionellen Theorie eines gehobenen Luftpakets (*Lifted Parcel Theory*), die in der Praxis für vereinfachte numerische Simulationen von hochreichender Feuchtkonvektion, Stabilitätsbetrachtungen (s. u.) und die Berechnung konvektiver Indizes (s. Kapitel 2.3) Anwendung findet, ist Adiabasie dennoch eine der zentralen Annahmen (z. B. Bjerknes, 1938; Holton, 2004). Insbesondere stellt diese Theorie lediglich ein eindimensionales Modell eines lokalen Luftpakets ohne horizontale Ausdehnung dar. Zu den weiteren Annahmen dieser Theorie zählt, dass (1) die feste Phase von Wasser nicht auftritt, d. h. keine Gefrierprozesse stattfinden, dass (2) der Effekt der Hydrometeore auf den Auftrieb vernachlässigbar ist und dass (3) der Druck des Luftpakets (Index P) zu jedem

Zeitpunkt dem der Umgebung (Index U) entspricht: $p_P = p_U = p$. Wie der Aufstieg des Luftpakets erzwungen wird, spielt für die Theorie eines gehobenen Luftpakets keine Rolle (vgl. Kapitel 1).
In der Praxis erfolgt die Berechnung einer trockenadiabatisch-pseudoadiabatischen Aufstiegskurve aus dem Anfangszustand eines Luftpakets, der durch den Zustand $(T_P, p_P, e_P) = (T_{P,A}, p_{P,A}, e_{P,A})$ gegeben ist (Abbildung 2.1). Die Aufstiegskurve stellt die Temperatur des Luftpakets T_P im jeweiligen Druckniveau dar. Zu Beginn steigt das Luftpaket trockenadiabatisch auf, d. h. mit $\theta_P = \theta_{P,A} = const.$ und $r_{v,P} = r_{v,P,A} = const.$, bis es eine Höhe erreicht, ab der Sättigung der Luft bezüglich des Wasserdampfs eintritt ($r_{v,P} = r^*_{v,P} = r_{v,s}$). Der Stern (*) kennzeichne Werte in diesem Niveau. Danach erfolgt der Aufstieg des Luftpakets pseudoadiabatisch, d. h. mit $\theta_{ps,P} = \theta^*_{ps,P} = const.$ Dazu wird zunächst $\theta^*_{ps,P}$ mit den Werten des Luftpakets T^*_P, p^* und $r^*_{v,P} = r_{v,s}$ mittels der Formel nach Bolton (1980), Gleichung (2.25), bestimmt. Anschließend muss die Temperatur des Pakets für die weiteren Druckniveaus des Aufstiegs $T_P(p < p^*)$ mittels Gleichung (2.25) mit den festen Werten für $\theta^*_{ps,P}$ und $r_{v,s}$ bestimmt werden. Es gilt dort $T_{\mathrm{HKN}}(T_P, e_s) = T_P$, sodass sich die Gleichung nicht analytisch nach T_P umformen lässt. Daher erfolgt die Bestimmung der Temperaturwerte iterativ. Entlang der Aufstiegskurve sind einige besondere Niveaus zu erwähnen:

Ausgangsniveau
Dieses Niveau charakterisiert die Höhe, in welcher der Anfangszustand des Luftpakets zu bestimmen ist. Die einfachste Methode nimmt an, dass das Luftpaket quasi vom Boden aufsteigt, sodass daher die Werte für $(T_{P,A}, p_{P,A}, e_{P,A})$ gewählt werden, die dem Boden am nächsten liegen. Diese entsprechen bei Radiosondenaufstiegen den Werten der ersten Messung[2]

[2] Da Radiosonden den Druck, die Temperatur und die relative Luftfeuchtigkeit RH messen, muss der Dampfdruck zunächst über $e = RH e_s(T)$ bestimmt werden.

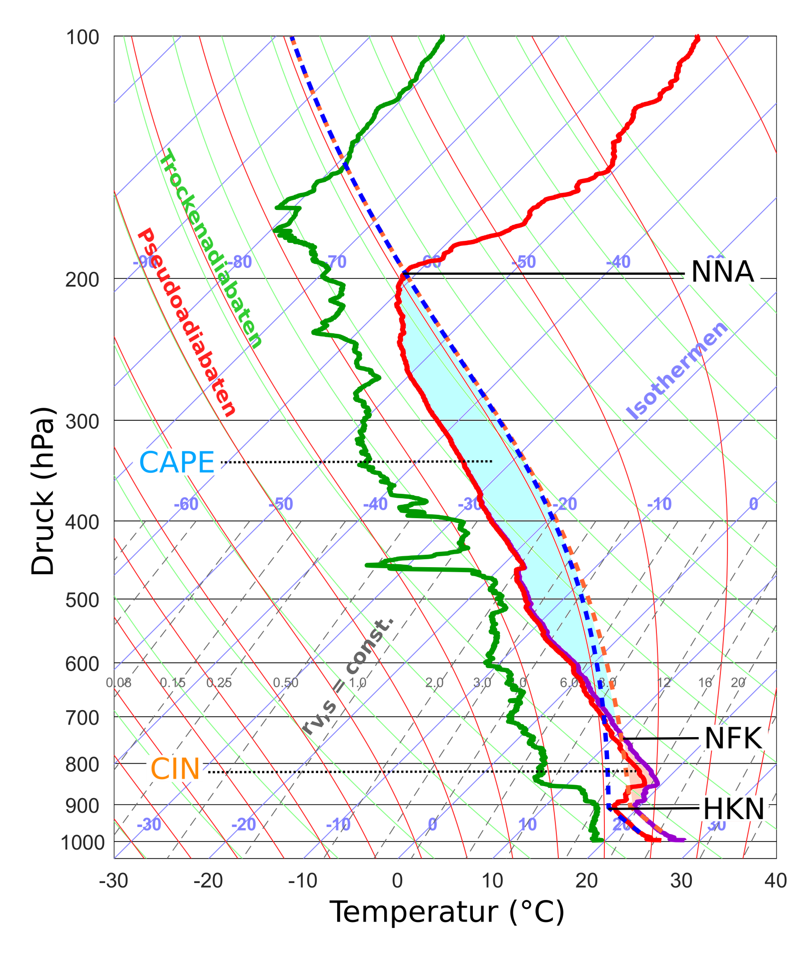


Abbildung 2.1: Thermodynamisches *Skew-T*-log*p*-Diagramm eines Radiosondenaufstiegs (Lindenberg, WMO: 10393; 11. Juni 2019, 12 UTC). Das Vertikalprofil der Temperatur der Umgebung T_U ist durch die rote Kurve, das des Taupunkts τ_U durch die grüne Kurve dargestellt. Die blau gestrichelte Kurve markiert den Verlauf der Temperatur T_P eines fiktiven, trockenadiabatisch-pseudoadiabatischen Aufstiegs eines Luftpakets. Die orangefarbene und die violette Kurve stellen jeweils die korrigierten virtuellen Temperaturprofile $T_{V,P}$ und $T_{V,U}$ dar. Die Werte für das Wasserdampfsättigungsmischungsverhältnis $r_{v,s}$ sind in g kg^{-1} angegeben. Trockenadiabaten repräsentieren $\theta(T,p) = const.$, (Sättigungs-)Pseudoadiabaten $\theta_{ps}(T,p,r_{v,s}) = const$. Nach Wilhelm et al. (2021).

oder den Werten der entsprechenden Bodenmessstation. In numerischen Modellen werden meist die Werte in der Mitte der untersten Modellschicht verwendet. Die zweite Methode bestimmt in einer sogenannten (trockenadiabatischen) Mischungsschicht (*Mixed Layer*; ML) mit einer Dicke von meist 50 bis 100 hPa über Grund dichtegewichtete Mittelwerte der Variablen. Craven et al. (2002) zeigten, dass ein auf diese Weise definiertes Luftpaket, welches in der Mitte der ML startet, für die Atmosphäre repräsentativer ist als eines, das auf den bodennahen Werten beruht. Die dritte Methode bestimmt in einer Schicht mit einer Dicke von meist 250 bis 300 hPa über Grund für verschiedene Niveaus oder Teilschichten die pseudopotentielle Temperatur (Manzato und Morgan, 2003). Dort, wo der höchste Wert gefunden wird, liegt das Ausgangsniveau für das – wörtlich übersetzt – instabilste Luftpaket (*Most Unstable*; MU).

Sättigungsniveau/Kondensationsniveau

Das zuvor mit dem Stern (*) gekennzeichnete Niveau, das die Höhe darstellt, in dem ein aufsteigendes Luftpaket zum ersten Mal Sättigung erreicht, wird im Fall erzwungener, dynamischer Hebung Hebungskondensationsniveau (HKN), im Fall thermischer Hebung Cumulus-Kondensationsniveau (KKN) genannt. Hier wird angenommen, dass instantan Kondensation und damit Wolkenbildung einsetzt.

Niveau freier Konvektion

Das unkorrigierte Niveau freier Konvektion (NFK) ist in der Höhe zu finden, in der die Aufstiegskurve die Kurve der Umgebungstemperatur T_U zum ersten Mal schneidet. Oberhalb des unkorrigierten NFK ist $T_P > T_U$. Freie Konvektion bezeichnet in der Theorie eines gehobenen Luftpakets den Umstand, dass

ein aufsteigendes Luftpaket einen positiven thermischen Auftrieb $B > 0$ relativ zu seiner Umgebung[3] besitzt, wobei aufgrund der Annahme $p_P = p_U$ keine Druckstörungen bzw. vertikale Gradienten in den Gleichungen (2.33) und (2.35) auftreten, die zur Vertikalbeschleunigung des Luftpakets beitragen könnten (z. B. Doswell und Markowski, 2004):

$$\left(\frac{\mathrm{D}w}{\mathrm{D}t}\right)_{LPT} \approx B_{LPT} = \frac{T_{V,P} - T_{V,U}}{T_{V,U}} g \; . \tag{2.36}$$

Aufgrund dieser Definition des thermischen Auftriebs ist klar, dass zur Bestimmung des korrigierten NFK die Temperaturprofile des Luftpakets und der Umgebung a posteriori über Gleichung (2.14) in Profile der entsprechenden virtuellen Temperatur T_V transformiert werden müssen (Abbildung 2.1; Doswell und Rasmussen, 1994). Das korrigierte NFK ist folglich in der Höhe zu finden, in der die $T_{V,P}$-Kurve die $T_{V,U}$-Kurve zum ersten Mal schneidet. Diese Korrektur berücksichtigt jedoch nur den Effekt von Wasserdampf auf die Dichte des Luftpakets (vgl. Kapitel 2.1.1). Der Effekt flüssigen Wassers oberhalb des HKN wird nicht korrigiert. Da das Luftpaket oberhalb des HKN gesättigt ist, ist die Korrektur der Aufstiegskurve T_P dort größer als die Korrektur des Temperaturprofils der meist ungesättigten Umgebung T_U. Doswell und Rasmussen (1994) merken an, dass die Entscheidung, wie man das Luftpaket und das Ausgangsniveau festlegt (s. o.), den Effekt dieser Korrektur überwiegen kann.

Niveau des neutralen Aufstiegs

Dieses Niveau charakterisiert die Höhe, ab der wieder $T_P < T_U$ bzw. $T_{V,P} < T_{V,U}$ gilt. Meist liegt es in der oberen Troposphäre, in der das Wasserdampfmischungsverhältnis r_v aufgrund der Abnahme des Sättigungsmischungsverhältnisses $r_{v,s}$ in der Berechnung der virtuellen Temperatur über

[3] Nicht in Bezug auf einen allgemeinen Grundzustand $\overline{p}(z)$.

Gleichung (2.14) einen vernachlässigbaren Beitrag liefert, sodass $T_V \approx T$ und $B_{LPT} \approx gT_P T_U^{-1}$ gilt (vgl. die Annäherung der beiden Aufstiegskurven in Abbildung 2.1). Am Niveau des neutralen Aufstiegs (NNA), an dem also wie am korrigierten NFK $B_{LPT} = 0$ ist, wird angenommen, dass instantan keine Kondensation mehr stattfindet, sodass dort die Wolkenobergrenze zu finden ist. Bei hochreichender Feuchtkonvektion liegt das NNA häufig in etwa auf der Höhe der Tropopause, an welcher der vertikale Temperaturgradient der Umgebung positiv wird.

Stabilitätsbetrachtungen

Aus den obigen Erläuterungen ist ersichtlich, dass der Aufstieg eines Luftpakets inhärent abhängig vom Zustand der Umgebung ist. Je nachdem, ob die atmosphärischen Bedingungen Konvektion physikalisch prinzipiell zulassen, spricht man allgemein von einer (statisch bzw. thermisch) stabil oder instabil (labil) geschichteten Umgebung. Instabil bedeutet dabei, dass eine kleine vertikale Auslenkung eines Luftvolumens zu einer sich selbst verstärkenden Störung des Gleichgewichtszustands, hier des hydrostatischen Gleichgewichts, anwächst. Neben der statischen Instabilität treten in der Atmosphäre eine Reihe weiterer Instabilitäten auf, die auf unterschiedlichen Raum- und Zeitskalen und unter bestimmten Strömungsverhältnissen für Horizontal- und Vertikalbewegungen eine Rolle spielen (z. B. Scherungs-, barokline, zentrifugale Instabilität; Markowski und Richardson, 2010; Vallis, 2017).
Aus dem Ersten Hauptsatz der Thermodynamik für adiabatische Zustandsänderungen aus Gleichung (2.7) und dem hydrostatischen Gleichgewicht aus Gleichung (2.31) folgt für einen horizontal homogenen Grundzustand einer trockenen Atmosphäre ($\overline{T} = \overline{T}(z)$ und $\overline{p} = \overline{p}(z)$):

$$-\Gamma_d \equiv \frac{\partial \overline{T}}{\partial z} = -\kappa g \frac{\overline{\rho}\,\overline{T}}{\overline{p}} \overset{(2.4)}{=} -\frac{g}{c_p} \approx -0{,}0098\,\mathrm{K\,m^{-1}} \,. \tag{2.37}$$

Die Temperatur eines solchen Grundzustands nimmt folglich linear mit der Höhe ab. Dabei heißt $-\Gamma_d$ trockenadiabatischer Temperaturgradient (z. B. Kraus, 2004). Nach einer etwas ausgedehnteren Rechnung (nicht gezeigt) erhält man für pseudoadiabatische Zustandsänderungen in einer gesättigten Atmosphäre aus Gleichung (2.23) analog mit $l_v(T) = l_{v,0} = const.$, der Näherung $p - e_s(T) \approx p$ und mit $l_{v,0} \gg R_v T$ den pseudoadiabatischen (irreversibel-feuchtadiabatischen) Temperaturgradienten:

$$-\Gamma_{ps}(z) \equiv \frac{\partial \overline{T}}{\partial z} = -\Gamma_d \frac{1 + \frac{l_{v,0} r_{v,s}(z)}{R_d \overline{T}(z)}}{\frac{\kappa}{\kappa_i(z)} + \kappa \frac{l_{v,0}}{R_v \overline{T}(z)} \frac{l_{v,0} r_{v,s}(z)}{R_d \overline{T}(z)}} \,. \tag{2.38}$$

Implizite und explizite Abhängigkeiten einiger Variablen von z sind dabei zur Verdeutlichung der Höhenabhängigkeit des Temperaturgradienten dargestellt. Die Ersetzung $\kappa_i \to \kappa_r$ in Gleichung (2.38) führt direkt auf den reversibel-feuchtadiabatischen Temperaturgradienten $-\Gamma_r$. Vernachlässigt man in Gleichung (2.38) den Beitrag von Flüssigwasser zur spezifischen Wärmekapazität ($\kappa_i \to \kappa$), so erhält man die etwas bekanntere Form des allgemeinen gesättigten, feuchtadiabatischen Temperaturgradienten

$$-\Gamma_s(z) \equiv \frac{\partial \overline{T}}{\partial z} = -\Gamma_d \frac{1 + \frac{l_{v,0} r_{v,s}(z)}{R_d \overline{T}(z)}}{1 + \kappa \frac{l_{v,0}}{R_v \overline{T}(z)} \frac{l_{v,0} r_{v,s}(z)}{R_d \overline{T}(z)}} \,. \tag{2.39}$$

Aus dem thermodynamischen Diagramm in Abbildung (2.1) kann man beispielsweise in der Höhe der 0 °C-Grenze für das Sättigungsmischungsverhältnis einen Wert von $r_{v,s} \approx 0{,}006\,\mathrm{kg\,kg^{-1}}$ ablesen, sodass eine Überschlagsrechnung in diesem Beispiel auf

$$-\Gamma_{ps}(z_{\vartheta=0\,^\circ\mathrm{C}}) \approx -0{,}0055\,\mathrm{K\,m^{-1}} \tag{2.40}$$

$$-\Gamma_s(z_{\vartheta=0\,^\circ\mathrm{C}}) \approx -0{,}0056\,\mathrm{K\,m^{-1}} \tag{2.41}$$

führt. Generell gilt in gesättigter Luft unter den in Kapitel 2.1.1 getroffenen Annahmen, dass das Mischungsverhältnis von Wasserdampf und Kondensat im reversiblen Prozess größer als das Sättigungsmischungsverhältnis von Wasserdampf im irreversiblen Prozess ist, d. h. es gilt $r_{w,0} \geq r_{v,s}$. Somit ist jederzeit $\kappa_r \leq \kappa_i \leq \kappa$ und $\Gamma_r \leq \Gamma_{ps} \leq \Gamma_s \leq \Gamma_d$. Die Unterschiede zwischen den verschiedenen feuchtadiabatischen Temperaturgradienten sind wie im obigen Beispiel jedoch meist gering (vgl. Markowski und Richardson, 2010). Der Wertebereich der feuchtadiabatischen Temperaturgradienten liegt etwa zwischen $-0{,}004\,\mathrm{K\,m^{-1}}$ und $-\Gamma_d$ und lässt sich mit dem Wissen über die ungefähre Höhenlage der jeweiligen Druckniveaus in Abbildung 2.1 gut erkennen.

Im Folgenden kennzeichne $-\gamma$ den vertikalen Temperaturgradienten einer beliebigen atmosphärischen Umgebung. Folgende differentielle Zustände statischer Schichtungsstabilität lassen sich unter der pseudoadiabatischen Annahme für gesättigte Luft unterscheiden:

- Absolute Stabilität: Trocken- oder feuchtadiabatisch gehobene Luftpakete kühlen sich in einer solchen Umgebung stärker ab als die Umgebung und steigen von selbst nicht weiter auf. Dies ist der Fall, wenn $\gamma < \Gamma_{ps}$ bzw. $\partial \theta_{ps,U}/\partial z > 0$ gilt.

- Absolute Instabilität: Trocken- oder feuchtadiabatisch gehobene Luftpakete kühlen sich in einer solchen Umgebung weniger stark ab als die Umgebung und können von selbst weiter aufsteigen. Dies ist der Fall, wenn $\gamma > \Gamma_d$ bzw. $\partial \theta_U/\partial z < 0$ gilt.

- Bedingte Instabilität: Trockenadiabatisch gehobene Luftpakete kühlen sich in einer solchen Umgebung stärker, feuchtadiabatisch gehobene Luftpakete weniger stark ab als die Umgebung. Dies ist der Fall, wenn $\Gamma_{ps} < \gamma < \Gamma_d$ bzw. $\partial \theta_{ps,U}/\partial z < 0$ und zugleich $\partial \theta_U/\partial z > 0$ gilt.

- Die Grenzfälle $\gamma = \Gamma_d$ ($\partial\theta_U/\partial z = 0$) und $\gamma = \Gamma_{ps}$ ($\partial\theta_{ps,U}/\partial z = 0$) bezeichnen eine trocken- bzw. feucht-neutrale Schichtung.

Der vertikale Temperaturgradient in der realen Atmosphäre liegt insbesondere während sommerlicher, konvektionsförderlicher Wetterlagen häufig über eine vertikal ausgedehnte Schicht der Troposphäre im Bereich der bedingten Instabilität. Recht häufig gibt es jedoch einerseits auch Schichten, die absolut stabil sind (wie z. B. zwischen 900 und 840 hPa in Abbildung 2.1). Andererseits liegen bisweilen Teile der Troposphäre teils im Grenzbereich trocken-neutraler Schichtung, beispielsweise bei mitteltroposphärischer Kaltluftadvektion bei gleichzeitigem Vorhandensein von Warmluft in den unteren Troposphärenschichten.
In manchen Situationen, in denen feuchte, warme Luft in den unteren Troposphärenschichten und trockene, kalte Luft darüber vorhanden ist, erhöht sich der Betrag des vertikalen Temperaturgradienten über diesen Bereich, wenn er als Ganzes gehoben wird. In den unteren Luftschichten setzt früher Kondensation ein als in den oberen, weswegen sie sich ab dem Zeitpunkt des Einsetzens der Kondensation weniger stark abkühlen. Dies führt zur Ausbildung der sogenannten potentiellen Instabilität in der Atmosphäre, sofern der Temperaturgradient in diesem Bereich nicht im Wertebereich absoluter Stabilität liegt, d. h. $\gamma > \Gamma_{ps}$ bzw. $\partial\theta_{ps,U}/\partial z < 0$ gilt (Rossby, 1932). Im Zusammenhang mit potentieller Instabilität kommt es auch häufig zu einer sogenannten abgehobenen Mischungsschicht (*Elevated Mixed Layer*; Carlson et al., 1983). Diese ist durch eine trockene Luftschicht in der mittleren Troposphäre (in Mitteleuropa in etwa im 700 hPa Niveau) gekennzeichnet, welche zuvor dynamisch oder orografisch trockenadiabatisch von bodennahen Niveaus ausgehend gehoben wurde. Werden solche Luftschichten in Regionen advehiert, in denen in der unteren Troposphäre warme, feuchte Luft vorzufinden ist, kann eine bedeutende potentielle Instabilität generiert werden (Lanicci und Warner, 1991). Dies war der Fall bei vielen schweren Hagelunwettern in Deutschland wie beispielsweise dem Münchner

Hagelunwetter am 12. Juli 1984 (Heimann und Kurz, 1985), dem Hagelsturm von Villingen-Schwenningen am 28. Juni 2006 (Noppel et al., 2010), dem Reutlinger Hagelunwetter am 28. Juli 2013 (Kunz et al., 2018) oder dem in München am 10. Juni 2019 (Wilhelm et al., 2021).

Bedingte Instabilität ermöglicht prinzipiell dann freie Konvektion, wenn ein mit Feuchte angereichertes Luftpaket aus der unteren Troposphäre zunächst trockenadiabatisch, nach Erreichen des HKN feuchtadiabatisch gehoben wird und dabei das NFK erreicht. Je kleiner jedoch γ und je trockener das Luftpaket ist, desto höher liegt das NFK und desto stärker muss der Hebungsantrieb sein, der das Luftpaket bis zum NFK anhebt. Diese Art bedingter Instabilität wird nach Normand (1931) als latente Instabilität bezeichnet und ist für die Entstehung hochreichender Feuchtkonvektion eine wichtige Voraussetzung (vgl. Kapitel 1; Groenemeijer, 2009; Mohr, 2013).
Zwei Maße zur Charakterisierung latenter Instabilität lassen sich aus einem thermodynamischen Diagramm mit einer berechneten Aufstiegskurve wie in Abbildung 2.1 unmittelbar ablesen: die konvektive verfügbare potentielle Energie (*Convective Available Potential Energy*, CAPE) und die konvektive Hemmung oder Sperre (*Convective Inhibition*, CIN). Die CAPE und die CIN sind abhängig von der Wahl der Methode für die Bestimmung des Ausgangsniveaus und des entsprechenden Luftpakets (s. o.). Sie sind als integrale Stabilitätsmaße wie folgt definiert:

$$\mathrm{CIN} = -\int_{z_B}^{\mathrm{NFK}} B_{LPT}\,\mathrm{d}z = -g\int_{z_B}^{\mathrm{NFK}} \frac{T_{V,P}-T_{V,U}}{T_{V,U}}\,\mathrm{d}z \qquad (2.42)$$

$$\mathrm{CAPE} = \int_{\mathrm{NFK}}^{\mathrm{NNA}} B_{LPT}\,\mathrm{d}z = g\int_{\mathrm{NFK}}^{\mathrm{NNA}} \frac{T_{V,P}-T_{V,U}}{T_{V,U}}\,\mathrm{d}z\,. \qquad (2.43)$$

Der Wertebereich der CAPE ist allgemein größer als der der CIN, d. h. ein Wert der CAPE von beispielsweise $100\,\mathrm{J\,kg^{-1}}$ ist ein relativ niedriger Wert, während CIN $= -100\,\mathrm{J\,kg^{-1}}$ schon eine bedeutsame konvektive Hemmung darstellt, die entweder durch Hebung überwunden oder durch verschiedene Prozesse im Lauf des Tages abgebaut werden muss (Markowski und Richardson, 2010). In Abbildung 2.1 ist für ein Mischungsschicht-Luftpaket $\mathrm{CAPE_{ML}} \approx 1900\,\mathrm{J\,kg^{-1}}$ (hellblaue Fläche) und $\mathrm{CIN_{ML}} \approx -100\,\mathrm{J\,kg^{-1}}$ (orange). In diesem Fall dauerte es bis zum Abend, bis sich starke Gewitter in der Umgebung entwickelten, weil der Hebungsantrieb tagsüber zu schwach war.

Latente Instabilität liegt demnach vor, wenn CAPE > 0 ist. Ob ein Luftpaket diese auch nutzen kann, hängt maßgeblich von seiner Feuchte und dem Vorhandensein eines genügend starken Hebungsantriebs ab, mit dessen Hilfe das Luftpaket die stabile Schicht überwinden kann (vgl. Doswell, 1987; Johns und Doswell, 1992). Je größer der Betrag der CIN ist, desto größer ist die Hemmung. Die CAPE und die CIN als kombinierte Maße von Instabilität und Feuchte müssen zur Vorhersage von hochreichender Konvektion demnach gemeinsam betrachtet werden und mit Indikatoren für verschiedene Hebungsprozesse ergänzt werden, um eine qualitative Aussage über das Auftreten der Konvektion treffen zu können.

Die Integration der mit der Vertikalgeschwindigkeit w multiplizierten Gleichung (2.36) über die Zeitspanne der freien Konvektion vom NFK bis zum NNA entspricht

$$\int_{w^2(t_{\mathrm{NFK}})}^{w^2(t_{\mathrm{NNA}})} \mathrm{d}w^2 = 2 \int_{z(t_{\mathrm{NFK}})}^{z(t_{\mathrm{NNA}})} B_{LPT}\,\mathrm{d}z\ . \tag{2.44}$$

Mit der Annahme, das $w(t_{\mathrm{NFK}})$ vernachlässigbar klein ist und w wegen der durchweg positiven Auftriebsbeschleunigung seinen größten Wert in der Höhe des NNA erreicht (Lohmann et al., 2016), folgt mit Gleichung (2.43)

die thermodynamische Grenzgeschwindigkeit als Obergrenze der Vertikalgeschwindigkeit durch Konvektion

$$w_{max} \equiv w(t_{\mathrm{NNA}}) \approx \sqrt{2\,\mathrm{CAPE}}\,. \tag{2.45}$$

Diese ist aufgrund der Vernachlässigung der (vertikalen) Druckstörungen, des Vorhandenseins von Hydrometeoren, die dem Auftrieb entgegenwirken, und des Einmischens trockener Umgebungsluft in der Theorie eines gehobenen Luftpakets besonders in der oberen Troposphäre deutlich größer als die tatsächliche Vertikalgeschwindigkeit (vgl. Gleichungen (2.34) und (2.35); Trapp, 2013). Häufig wird w_{max} in Grafiken verwendet (wie beispielsweise in Kapitel 5), da wegen des Wurzelziehens der Wertebereich der CAPE gestaucht wird.

2.2 Gewittersysteme und ihr Lebenszyklus

Nachdem die dynamischen und thermodynamischen Grundlagen erklärt wurden, beschäftigt sich dieses Kapitel mit den verschiedenen Organisationsformen konvektiver Zellen. Im Folgenden wird zunächst das Modell des Lebenszyklus einer idealisierten Einzelzelle ausführlich beschrieben, da dieses essentiell für das Verständnis der weiteren Organisationsformen ist (Kapitel 2.2.2 bis 2.2.4). Um eine Verbindung zu den thermodynamischen und dynamischen Umgebungsbedingungen verschiedener konvektiver Systeme herzustellen, werden speziell die CAPE sowie die vertikale Windscherung, ausgedrückt durch den Betrag der Differenz der horizontalen Windvektoren in Bodennähe und in 6 km über Grund (*Deep Layer Shear*, DLS), in den jeweiligen Abschnitten diskutiert. Eine Vorstellung weiterer Umgebungsvariablen und konvektiver Indizes folgt in Kapitel 2.3.

2.2.1 Isolierte Konvektion – Einzelzellen

Grundlegende Beschreibungen des Lebenszyklus einer einzelnen konvektiven Zelle gehen auf die Analysen von Byers und Braham (1948) im Rahmen des sogenannten *Thunderstorm Projects* zurück, das im Jahr 1947 in Ohio durchgeführt wurde[4]. Anhand einer Kombination von Beobachtungsdaten aus Flugzeugmessungen und aus bodengebundenen Radarmessungen gelang es erstmals, organisierte Strukturen einer konvektiven Zelle, insbesondere deren Auf- und Abwindbereiche, zu identifizieren (z. B. Doswell, 2007). Auf dieser Basis entstand das erste konzeptionelle Modell des Lebenszyklus einer Einzelzelle, welches Doswell (1985) mit den darauf aufbauenden Erkenntnissen aus weiteren knapp 40 Jahren experimenteller, theoretischer und numerischer Forschung ergänzte. Dieses Modell beschreibt den Lebenszyklus als Abfolge von drei separaten Entwicklungsstufen: 1) Wachstumsstadium, in Anlehnung an das Englische (*[Towering] Cumulus Stage*) meist Cumulusstadium genannt, 2) Reifestadium (*Mature Stage*) und 3) Dissipationsstadium (*Dissipation Stage*). Diese Einteilung basiert maßgeblich auf der Veränderung dynamischer und mikrophysikalischer Eigenschaften der Zellen im Verlauf des Lebenszyklus und soll im Folgenden in Anlehnung an Doswell (1985), Markowski und Richardson (2010) und Trapp (2013) kurz dargestellt werden.

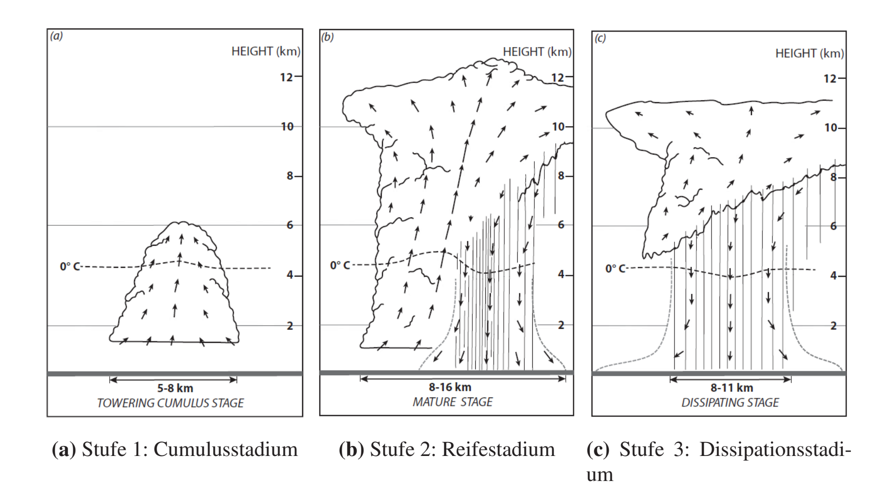
1) Cumulusstadium

Ähnlich der Vorstellung eines aufsteigenden Luftpakets ohne räumliche Ausdehnung in der Theorie eines gehobenen Luftpakets (vgl. Kapitel 2.1.2) beginnt der Lebenszyklus einer Einzelzelle mit einem adiabatisch zum NFK gehobenen oder von selbst aufsteigenden Luftpaket mit endlicher räumlicher Ausdehnung. Durch das Überschreiten des HKN setzt Kondensation durch heterogene Nukleation ein, sodass sich eine Cumulus-Wolke bildet. Oberhalb

[4] `https://www.weather.gov/iln/ThunderstormProject`

des NFK entwickelt sich im Inneren der Wolke aufgrund des positiven Auftriebs ($B_{LPT} > 0$ in Gleichung (2.36)) eine positive Vertikalbeschleunigung, welche zu einem sich vertikal intensivierenden Aufwindbereich (*Updraft*) führt (Abbildung 2.2a). Infolge dieser Vertikalbewegung entsteht unterhalb des Luftpakets eine konvergente horizontale Strömung. Die sich entwickelnde Konvektionszelle wird dadurch mit weiterer Feuchtigkeit aus der Umgebung versorgt, sodass durch das Freisetzen weiterer latenter Wärme oberhalb des HKN der Aufwindbereich der Zelle gestärkt werden kann. Rasch erreicht die Cumulus-Wolke eine vertikale Mächtigkeit von mehreren Kilometern (Cumulus congestus bzw. *Towering Cumulus*) und einen horizontalen Durchmesser von etwa 5 – 8 km.

Während des Übergangs vom Cumulus- ins Reifestadium ist die Wolke zu einem Cumulonimbus angewachsen, dessen Obergrenze bereits das NNA erreicht. Die Wolkentröpfchen wachsen durch das Kollidieren und Zusammenfließen mehrerer Tröpfchen an (Koaleszenz). Aufgrund der sinkenden Temperaturen in der oberen Troposphäre bilden sich neben unterkühlten Wolkentröpfchen zunehmend auch Wolkeneispartikel, die den Wegener-Bergeron-Findeisen-Prozess initiieren können, der im weiteren Verlauf zu einem beschleunigten Wachstum der Eispartikel auf Kosten der unterkühlten Wassertröpfchen führt. Details zur komplexen Wolkenmikrophysik finden sich beispielsweise in Seinfeld und Pandis (2006), Wang (2013) oder Lohmann et al. (2016). Wenn die Hydrometeore ein Gewicht und damit einen negativen Auftrieb erreichen, der den positiven Auftrieb im Aufwindbereich kompensiert, beginnen sie zu fallen.



(a) Stufe 1: Cumulusstadium **(b)** Stufe 2: Reifestadium **(c)** Stufe 3: Dissipationsstadium

Abbildung 2.2: Dreistufiges, konzeptionelles Modell des Lebenszyklus einer idealen Einzelzelle nach Byers und Braham (1948) und Doswell (1985), übernommen aus Trapp (2013). © Cambridge University Press (genehmigter Nachdruck).

2) Reifestadium

Doswell (1985) legt dar, dass unterschiedliche Definitionen für den Beginn des Reifestadiums geeignet und gebräuchlich sind. Häufig gilt als Beginn des Reifestadiums der Zeitpunkt, zu dem der erste Niederschlag den Boden erreicht. Aus dynamischer Perspektive ist für das Reifestadium die Entwicklung einer Region mit absinkender Luft, initiiert durch fallende Hydrometeore, (negative Vertikalbeschleunigung; $B_{LPT} < 0$; *Downdraft*) essentiell, was meist schon rund 10 min, bevor der erste Niederschlag den Boden erreicht, der Fall ist. Zur weiteren Entwicklung des Abwindbereichs trägt nicht nur der fallende Niederschlag bei, sondern auch das Einmischen trockener Luft aus der direkten Umgebung der Wolke, welches die Evaporation kleiner Wassertröpfchen fördert (vgl. Kapitel 2.1.2). Aufgrund der eher geringen horizontalen Ausdehnung einer Einzelzelle spielen die durch den Aufstieg induzierten vertikalen Druckstörungen für die Dynamik eine untergeordnete Rolle.

Der Wegener-Bergeron-Findeisen-Prozess und weitere niederschlagsbildende Prozesse laufen im oberen Bereich der Wolke am effektivsten ab. Durch das Erreichen des NNA bedingt die Massenkontinuität gemäß Gleichung (2.28) dort eine horizontal divergente Strömung, welche die Ausbildung horizontal ausgedehnter Eiswolken, den Amboss, initiiert und aufrechterhält. In Kombination mit der höheren mittleren Strömungsgeschwindigkeit in der oberen Troposphäre führt dies zu einer (leichten) Asymmetrie der Cumulonimbuswolke. Die stetig wachsenden Hydrometeore fallen daher als intensive Niederschläge häufig etwas stromab versetzt zum Aufwindbereich Richtung Erdboden (Abbildung 2.2b). Dabei findet ein weiteres Einmischen der trockeneren Umgebungsluft statt, welches den Abwind verstärkt. Schnell erreicht der Abwind mit den intensiven Niederschlägen den Boden und strömt näherungsweise horizontal isotrop in den bodennahen Luftschichten auseinander. Diese Diffluenz (*Outflow*) vergleichsweise kühler Luft (*Cold Pool*) führt durch Geschwindigkeitskonvergenz zur Ausbildung einer Böenfront im Grenzbereich zur bodennahen Umgebungsluft.
Gleichzeitig beginnt damit der Übergang in das Dissipationsstadium: Der fallende Niederschlag und die durch den Abwindbereich induzierte bodennahe Diffluenz führen dazu, dass die ursprüngliche bodennahe Konvergenz feucht-warmer Luft unterhalb des Aufwindbereichs abgeschwächt und schließlich aufgehoben wird, sodass aufgrund der Kontinuitätsgleichung (2.28) der Massenfluss in den Aufwindbereich versiegt. Nichtsdestoweniger kommt es in der verbleibenden Cumulonimbuswolke zunächst weiterhin zum auftriebsbedingten Aufsteigen, weiterer Kondensation und Niederschlagsbildung.

3) Dissipationsstadium

Durch den fehlenden Nachschub an feucht-warmer Luft nimmt der Auftrieb im Aufwindbereich vom Boden her immer weiter ab, sodass sich die Aufwinde weiter abschwächen, bis die gesamte Zelle nur noch aus einem großen

Abwindbereich und fallendem Niederschlag besteht (Abbildung 2.2c). Die Böenfront breitet sich in den untersten Schichten der Troposphäre weiter horizontal aus und verliert durch die Abnahme der Intensität des Abwindbereichs und des Niederschlags an Stärke. Die Wolke wird nach und nach von unten her aufgelöst und es bleiben Überreste des Ambosses in der oberen Troposphäre übrig, welche in der Folge evaporieren bzw. sublimieren.

Folgende besondere Merkmale treten während des Lebenszyklus einer Einzelzelle auf:

- Blitze: Je stärker die Aufwinde einer konvektiven Zelle sind, desto häufiger treten Blitze auf, besonders im Reifestadium. Die sich bildenden Wassertröpfchen und Eispartikel wachsen und formen mit der Zeit größere Hydrometeore, insbesondere Graupel, der ein Eisteilchen mit einer geringeren Dichte und einem Durchmesser von bis zu 5 mm darstellt. Kollidieren kleine Eispartikel mit dem größeren Graupel, kommt es zur Ionisierung der Stoßpartner. Wegen seines wachsenden Gewichts fällt der nun negativ geladene Graupel in niedrigere Bereiche der Wolke, während die leichten, positiv geladenen Eispartikel durch die Aufwinde in den oberen Bereich der Wolke transportiert werden. Im unteren Bereich der Wolke beginnen die Graupelkörner – zumindest bei sommerlicher Konvektion – zu schmelzen und laden sich dabei leicht positiv auf. Eine Ladungstrennung ist erfolgt und die Wolke kann als ein elektrischer Tripol angesehen werden (Rakov und Uman, 2003). Dies ist jedoch nur eine von mehreren Modellvorstellungen, die derzeit diskutiert werden. Bei entsprechend großer Ladungstrennung kommt es in der Folge zu Entladungen in Form von Blitzen innerhalb einer Wolke (*Intra-Cloud Lightning*), zwischen zwei benachbarten Wolken (*Cloud-to-Cloud Lightning*) oder zwischen der Wolke und dem

Boden (*Cloud-to-Ground Lightning*). Im Dissipationsstadium reduziert sich die Ladungstrennung aufgrund der nachlassenden Aufwinde und es werden kaum noch Blitze beobachtet.

- Konvektives Überschießen (*Overshooting Top*): Je stärker der Aufwindbereich einer konvektiven Zelle, desto eher reichen die Vertikalgeschwindigkeiten im oberen Bereich der Zelle aus, dass der Aufwind das NNA in einem gewissen Maß durchbrechen kann. Oberhalb des Ambosses sind dann Wolken zu finden, die eine kuppelartige Form aufweisen. Aufgrund der sehr stabilen Schichtung in diesem Bereich stratosphärischer Luft beginnen die übergeschossenen Luft- und Wasserteilchen rasch wieder abzusinken (z. B. Doswell, 1985). Das konvektive Überschießen ist bei Multi- und Superzellen (Kapitel 2.2.2 und 2.2.3) meist stärker ausgeprägt als bei Einzelzellen; in der Praxis werden automatische Detektionen des konvektiven Überschießens in Satellitenbildern daher beispielsweise als Proxy für Hagel verwendet (z. B. Bedka, 2011; Punge et al., 2017).

Einzelzellen entstehen vor allem, wenn eine geringe Windscherung mit etwa DLS $\leq 10\,\mathrm{m\,s^{-1}}$ vorherrscht (Markowski und Richardson, 2010; Trapp, 2013). Der Wertebereich der CAPE hingegen ist weniger entscheidend, wobei die meisten Einzelzellen bei niedrigen bis moderaten Werten bis etwa CAPE $= 1\,000\,\mathrm{J\,kg^{-1}}$ auftreten. Typischerweise herrschen solche Bedingungen während synoptisch gradientschwacher Wetterlagen vor, bei denen Konvektion vor allem durch den Tagesgang der Temperatur und Feuchte in der atmosphärischen Grenzschicht infolge von solarer Einstrahlung bestimmt wird. Häufig dienen dann orografisch bedingte Hebung oder lokale horizontale Strömungskonvergenzen als Auslösemechanismus. Mit einer typischen vertikalen Ausdehnung von $H = 10\,\mathrm{km}$ und Vertikalgeschwindigkeiten im

Auf- und Abwindbereich von etwa $W = 5 - 10\,\mathrm{m\,s^{-1}}$ erhält man für die typische Lebensdauer einer Einzelzelle etwa

$$T_Z \approx 2\,\frac{H}{W} \approx 30 - 60\,\mathrm{min}\,, \qquad (2.46)$$

welche das einmalige Durchlaufen eines Luftpartikels durch einen Auf- und Abwindbereich charakterisiert.

2.2.2 Multizelluläre Konvektion

Eine Multizelle setzt sich – wie der Name bereits verrät – aus mehreren Einzelzellen zusammen, die dynamisch miteinander interagieren. Die Zellen sind dabei in unterschiedlichen Stadien ihrer Entwicklung, wie in Abbildung 2.3 illustriert ist: Hier befindet sich in der oberen Abbildung die als erste aufgetretene Zelle 1 bereits im Dissipationsstadium, Zelle 2 am Ende des Reifestadiums, Zelle 3 am Beginn des Reifestadiums und Zelle 4 im Cumulusstadium. Eine solche organisierte Entwicklung ist nur möglich, wenn die Troposphäre ausreichend labil geschichtet ist und eine moderate vertikale Windscherung vorliegt (DLS $\approx 10 - 20\,\mathrm{m\,s^{-1}}$). Eine besonders geeignete Kenngröße, mittels derer das Auftreten multizellulärer Konvektion gut charakterisiert werden kann, ist die *Bulk Richardson Number* (BRN). Diese verknüpft die potentielle Energie in der Umgebung, charakterisiert durch die CAPE, mit der kinetischen Energie in der Umgebung, charakterisiert durch ein Maß für die mittlere vertikale Scherung in einer hochreichenden vertikalen Schicht:

$$\mathrm{BRN} = \frac{\mathrm{CAPE}}{0{,}5\,|\Delta\overline{\mathbf{u}}|^2}\,. \qquad (2.47)$$

Darin lässt sich $\Delta\overline{\mathbf{u}}$ beispielsweise durch die Differenz der horizontalen Windvektoren bestimmen, die den mittleren Wind in der Umgebung zwischen 0 und 6 km über Grund und den mittleren Wind in einer bodennahen Schicht wie z. B. zwischen 0 und 0,5 km über Grund angeben (Weisman und Klemp, 1982; Markowski und Richardson, 2010):

$$\Delta\overline{\mathbf{u}} = \overline{\mathbf{u}}_{0-6\mathrm{km}} - \overline{\mathbf{u}}_{0-0,5\mathrm{km}} \,. \tag{2.48}$$

Ist etwa BRN > 35, so herrschen für die Entwicklung von Multizellen förderliche Umgebungsbedingungen vor (Trapp, 2013).
Die Dynamik einer idealisierten Multizelle kann wie folgt erläutert werden: Durch die ausgeprägte Böenfront einer (Multi-)Zelle infolge starker Abwinde kommt es stromab des vertikalen Scherungsvektors $\mathbf{S} = \partial \mathbf{v}/\partial z$ durch die Interaktion mit dem bodennahen horizontalen Vorticityfeld zur Hebung warmer Luft aus der Umgebung, welche zur Ausbildung eines neuen, vorgelagerten Aufwindbereichs führt (Abbildung 2.3, oben). In kurzer Zeit entwickelt sich eine neue Zelle, in der niederschlagsbildende Prozesse einsetzen (Zelle 4 in Abbildung 2.3, Mitte). Gleichzeitig erhält die vorherige Zelle (Zelle 3) dadurch immer weniger Nachschub an feucht-warmer Luft und geht auf das Ende ihres Reifestadiums zu. Ihr Niederschlagsbereich verschmilzt mit den schwächer werdenden Niederschlägen der Zellen 1 und 2, ihre Abwinde erreichen den Boden und stärken die bereits vorhandene Böenfront. Kurze Zeit später befindet sich Zelle 4 im Übergang in das Reifestadium und die Böenfront induziert erneut die Entwicklung eines neuen Aufwindbereichs (Zelle 5; Abbildung 2.3, unten).
Ohne oder bei geringer vertikaler Windscherung reicht die durch die Böenfront induzierte Hebung meist nicht aus, damit die aufsteigende Luft stromab des Scherungsvektors das NFK erreicht. Erst das durch die Geschwindigkeitsscherung hervorgerufene Vorticityfeld, dessen Rotationsachse senkrecht zum Scherungsvektor aus der Zeichenebene in Abbildung 2.3 hinaus steht,

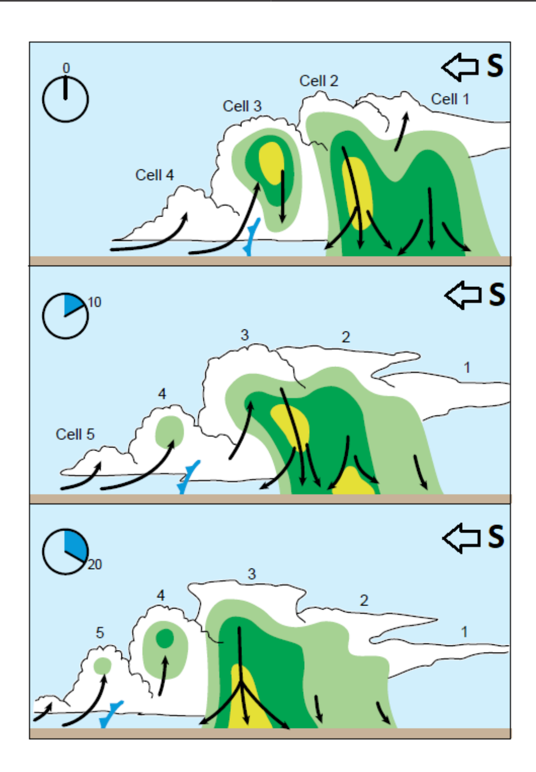


Abbildung 2.3: Schematische Darstellung einer Multizelle nach Doswell (1985), übernommen aus Markowski und Richardson (2010) und leicht modifiziert. Dünne Pfeile charakterisieren Auf- und Abwindbereiche. Niederschlag ist durch die grün (schwach bis mäßig) und gelb (stark) eingefärbten Bereiche dargestellt. Die Lage der Böenfront im Vorfeld des Bereichs der intensivsten Niederschläge in Richtung des vertikalen Scherungsvektors **S** (hier vereinfacht als richtungskonstant angenommen) ist durch die kleine kaltfrontartige Linie gekennzeichnet. Zusätzlich ist links die verstrichene Zeit in Minuten angegeben. © John Wiley & Sons (genehmigter Nachdruck).

ermöglicht dies (Fovell und Tan, 1998; Lin et al., 1998). Die kalte, ausströmende Luft weist ihrerseits eine negative vertikale Geschwindigkeitsscherung auf, die ein entgegengesetzt gerichtetes Vorticityfeld induziert (Rotationsachse senkrecht zum Scherungsvektor in die Zeichenebene hinein). An der Böenfront addieren sich somit die Beiträge beider Vorticityfelder, sodass die Hebung stromab des Scherungsvektors deutlich stärker ausfällt als bei

kaum vorhandener Windscherung. Durch diese Vertikalbewegungen ausgelöste Schwerewellen können während des Cumulusstadiums der Zellen den Aufwindbereich weiter verstärken (z. B. Lin und Joyce, 2001).
Weiterhin spielen Faktoren wie das Maß an Richtungsscherung der mittleren Winde, mesoskalige Inhomogenitäten der Umgebungsbedingungen, die durch die Orografie, Konvergenzen oder unterschiedliche Landoberflächen hervorgerufen werden, sowie großskalige Hebungsprozesse eine Rolle für die genaue, real ablaufende Dynamik einer Multizelle. Die Hebung entlang der Böenfront ist abhängig von der bodennahen Strömung, die lokal- und mesoskalig deutlich von der mittleren Windrichtung in der unteren und mittleren Troposphäre abweichen kann, welche in etwa mit der Zugrichtung der Einzelzellen assoziiert werden kann. Das Bild der Vorticityfelder muss daher auf eine horizontale Ebene erweitert werden. Die beiden Vorticityfelder addieren sich am effektivsten, wenn der bodennahe Wind senkrecht in Richtung zur Böenfront steht. Unter der Annahme eines konstanten Windfelds ist dort die Hebung am größten, sodass sich die nächste Zelle in diesem Bereich bildet (Abbildung 2.4). Im Allgemeinen entsteht die nächste Zelle, wo die Konvergenz des bodennahes Windes mit der Böenfront am stärksten ausgeprägt ist.
Der Bewegungsvektor einer Multizelle $\mathbf{c}_Z$ ist als eine effektive Verlagerung des Komplexes zu verstehen[5]. Diese ergibt sich durch die Vektoraddition der Verlagerung der Einzelzellen mit der mittleren Strömung $\overline{\mathbf{u}}$ der Schichten, über die sich die Wolken erstrecken, und der Entwicklungsrichtung des Systems $\mathbf{c}_{neu}$ (auch Propagations- oder Zellneubildungsvektor genannt), die durch die Position der sich neu entwickelnden Zellen vorgegeben

[5] Der Index Z kennzeichnet in der vorliegenden Arbeit Attribute einer konvektiven Zelle, die in mehreren Kapiteln vorkommen, und dient zur besseren Abgrenzung gegenüber den Umgebungsvariablen.

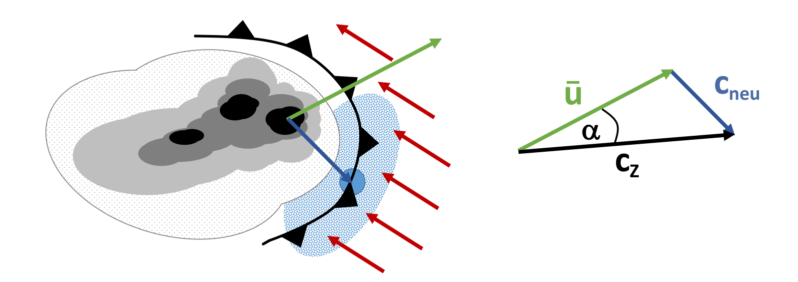


Abbildung 2.4: Schematische Darstellung zur Verlagerung einer Multizelle nach Houze (1993). Die Multizelle ist als ein zusammenhängendes Gebiet des Radarreflektivitätsfaktors durch die graue Umrandung dargestellt. Niedrige Reflektivitäten sind gepunktet, hohe in gefüllten Grautönen abgestuft dargestellt. Rote Pfeile repräsentieren das bodennahe Windfeld. Der blau schraffierte Bereich kennzeichnet den Bereich der höchsten Konvergenz. Der blaue Kreis markiert den Ort einer Neubildung einer Zelle, der für das Vektordiagramm rechts beispielhaft verwendet wird.

wird (vgl. Abbildung 2.4):

$$\mathbf{c}_Z = \overline{\mathbf{u}} + \mathbf{c}_{neu} \ . \tag{2.49}$$

Der Winkel α zwischen der mittleren Windrichtung $\overline{\mathbf{u}}$ und dem Verlagerungsvektor $\mathbf{c}_Z$ kann nach Marwitz (1972b) in Einzelfällen über 50° betragen. Dies ist insbesondere dann der Fall, wenn die Verlagerung der einzelnen Zellen doch etwas von der mittleren Windrichtung abweicht, wie es bei starker Richtungsscherung in der unteren und mittleren Troposphäre vorkommt. Durch die Bildung neuer Zellen kann die Lebensdauer des gesamten konvektiven Systems deutlich länger als bei einer Einzelzelle sein. Die durch die Böenfront neu initiierte Zelle kann jeweils relativ schnell dafür sorgen, dass die zuvor gebildete Zelle vom Einströmbereich der warm-feuchten Umgebungsluft abgeschnitten wird. Wie in Abbildung 2.3 näherungsweise dargestellt, beträgt die Zeitspanne zwischen der Bildung zweier Zellen häufig nur etwa 15 min (Fovell und Dailey, 1995). Durch die insgesamt längere Lebensdauer und die dynamische Interaktion der einzelnen Zellen in einer Multizelle

können allgemein intensivere und länger andauernde Niederschläge als bei Einzelzellen auftreten. Zudem ist die Böenfront in der Regel stärker ausgeprägt, sodass höhere Böengeschwindigkeiten gemessen werden. Auch die Bildung von meist kleinem bis mittelgroßem Hagel (bis etwa 5 cm) ist aufgrund längerer Trajektorien der Hydrometeore als bei einer Einzelzelle möglich (z. B. Browning, 1977).

2.2.3 Isolierte Konvektion – Superzellen

Superzellen treten in Europa deutlich seltener auf als kurzlebige Einzelzellen oder multizelluläre Systeme. Sie können aufgrund ihres hohen Grades an Organisation über mehrere Stunden bestehen, Zugbahnen mit einer Länge von einigen hundert Kilometern erreichen und große Schäden durch konvektive Starkwindböen, Tornados und großen Hagel verursachen. Im Gegensatz zu einer Multizelle, deren Lebensdauer und Verlagerung eng mit der Entwicklung neuer konvektiver Zellen stromab des Systems verbunden ist, handelt es sich bei einer Superzelle um isolierte Konvektion, deren Entwicklung durch die Dynamik eines starken, rotierenden Aufwindbereichs bestimmt ist (einer sogenannten Mesozyklone; z. B. Lemon und Doswell, 1979). Indem permanent warm-feuchte Luftmassen in diesen Aufwindbereich advehiert werden und niederschlagsinduzierte Abwindbereiche aufgrund einer ausgeprägten vertikalen Windscherung räumlich vom Aufwindbereich separiert entstehen, kann sich eine Superzelle über viele Stunden selbst erhalten. Das charakteristische Erscheinungsbild einer Superzelle besteht dabei aus der rotierenden Mesozyklone und zwei flankierenden Abwindbereichen, einem an der vorderen Flanke des Aufwindbereichs (*Forward-Flank Downdraft*) und einem an der rückseitigen Flanke (*Rear-Flank Downdraft*; Abbildung 2.5).

In Abhängigkeit vom genauen vertikalen Windprofil in der Umgebung können Superzellen im Laufe ihres Lebenszyklus unterschiedlich ausgeprägte Charakteristika entwickeln. Eine mögliche, grobe Klassifikation orientiert sich am vertikalen Windprofil in der mittleren und oberen Troposphäre,

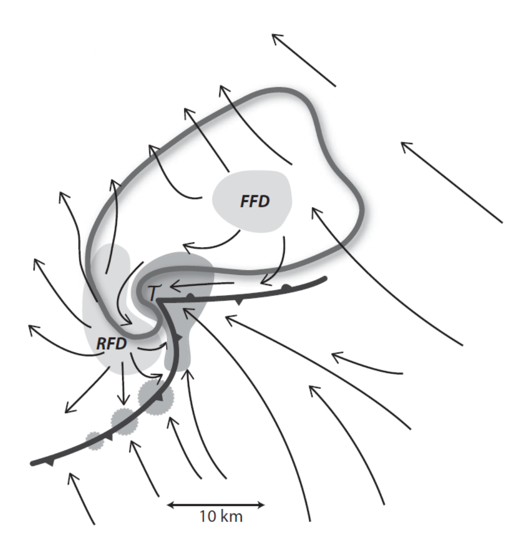


Abbildung 2.5: Aufsicht auf eine klassische Superzelle zum Höhepunkt ihres Reifestadiums nach der Entwicklung einer Mesozyklone, hier durch ein *T* angedeutet (Horizontalschnitt in der unteren Troposphäre; nach Lemon und Doswell (1979), übernommen aus Trapp (2013)). Die graue Kontur kennzeichnet die Grenze starken Niederschlags, die dunkelgrauen Flächen stellen Aufwind-, die hellgrauen Flächen Abwindbereiche dar (FFD: *Forward-Flank Downdraft*; RFD: *Rear-Flank Downdraft*). Das sturm-relative Windfeld ist durch die Pfeile dargestellt, dessen Konvergenzen durch die dunkelgraue Frontlinie. © Cambridge University Press (genehmigter Nachdruck).

d. h. im Bereich, wo die niederschlagsbildenden Prozesse besonders effektiv ablaufen. Klassische Superzellen bei moderaten sturm-relativen Winden in den höheren Schichten weisen im Radarbild das größte Niederschlagsgebiet im vorderseitigen Abwindbereich auf. Der intensivste Niederschlag fällt im rückseitigen Abwindbereich und nimmt aufgrund der Interaktion mit dem rotierenden Aufwindbereich eine hakenförmige Struktur an (*Hook Echo*). Der Aufwindbereich selbst ist im Radarbild durch niedrige Niederschlagssignale gekennzeichnet (*Bounded Weak Echo Region*; z. B. Browning, 1965; Marwitz, 1972a).

Bei starken sturm-relativen Winden in den höheren Schichten werden die Hydrometeore rasch horizontal vom Aufwindbereich weg verfrachtet, sodass sie nicht allzu stark anwachsen und als Niederschlag (sturm-relativ) vor allem stromab des Aufwindbereichs im vorderseitigen Abwindbereich Richtung Erdboden fallen. Die Niederschlagsintensität ist meist moderat, bisweilen wird jedoch auch größerer Hagel beobachtet (*Low Precipitation Supercell*). Im Gegensatz dazu haben bei schwachen sturm-relativen Winden in der oberen Troposphäre die Hydrometeore eine größere Verweildauer im Aufwindbereich und können daher zu größeren Regentropfen und Hagelkörnern anwachsen. Der Niederschlag fällt anschließend im rückseitigen Abwindbereich, der näher am Aufwindbereich und der rotierenden Mesozyklone liegt. Größere Teile des Niederschlagsbereichs können in die Rotation der Mesozyklone mit eingebunden werden und so die Rotation insbesondere in den unteren Schichten verstärken. Gleichzeitig können Teile des Niederschlagsbereichs die Aufwinde schwächen oder dem Einströmen warm-feuchter Luftmassen in den Aufwindbereich entgegenwirken. Die Niederschlagsintensität ist in diesen Superzellen meist sehr hoch, das Hakenecho stark ausgeprägt und die Wahrscheinlichkeit für großen Hagel und Tornadobildung erhöht (*High Precipitation Supercell*). Die genauen Details der hier vereinfacht beschriebenen Prozesse sind allerdings sehr sensitiv in Bezug auf das vertikale Windprofil über die gesamte Höhe der Troposphäre. Dieses bestimmt folglich maßgeblich die Intensität und die Lebensdauer einer Superzelle.

Förderliche Umgebungsbedingungen für die Entwicklung einer Superzelle sind neben einer hohen vertikalen Windscherung zwischen den bodennahen und höheren Luftschichten (DLS $> 18\,\mathrm{m\,s^{-1}}$; vor allem Richtungsscherung) ausgedehnte Feuchtefelder in der unteren Troposphäre und ein ausgeprägter Grenzschichtstrahlstrom (*Low Level Jet*), der effektiv feucht-warme Luft in den Aufwindbereich einströmen lässt (z. B. Johns und Doswell, 1992). Großräumige Hebung beispielsweise auf der Vorderseite eines Höhentrogs kann die potentielle Instabilität erhöhen. Besonders im Zusammenspiel mit einer

abgehobenen Mischungsschicht führt dies zu hohen Werten der CAPE, welche für die Entwicklung von starken Aufwinden und damit von Superzellen förderlich sind (vgl. Kapitel 2.1.2).

Entstehung einer rotierenden Mesozyklone

Entscheidend für den Lebenszyklus einer Superzelle ist die Genese einer rotierenden Mesozyklone (z. B. Davies-Jones, 1984; Klemp, 1987; Markowski und Richardson, 2010; Trapp, 2013). Dieser dynamisch komplexe Prozess setzt etwa während des Übergangs vom Cumulus- in das Reifestadium der Zelle ein, wenn sich bereits ein ausgeprägter Aufwindbereich entwickelt hat. Ausgangspunkt für die theoretische Betrachtung ist die Impulsgleichung (2.27). Reibungseffekte seien ausgenommen, die Coriolis- und Zentrifugalbeschleunigung durch die Erdrotation aufgrund der betrachteten räumlichen Skala ($L = 10\,\mathrm{km}$) vernachlässigbar und die Schwerebeschleunigung wie schon in Kapitel 2.1.2 zu $\mathbf{g} \approx g\,\mathbf{e}_z$ vereinfacht. Die Anwendung der Rotation auf Gleichung (2.27) und anschließende Projektion auf den vertikalen Einheitsvektor führt auf die vertikale Vorticitygleichung

$$\frac{\mathrm{D}\zeta}{\mathrm{D}t} = -\zeta \nabla \cdot \mathbf{u} + \boldsymbol{\omega}_h \cdot \nabla w - \mathbf{e}_z \cdot (\nabla \alpha \times \nabla p) \ , \tag{2.50}$$

in der ζ für die vertikale und $\boldsymbol{\omega}_h$ für die horizontale Komponente der relativen Vorticity $\boldsymbol{\omega} = \nabla \times \mathbf{v}$ steht. Der Beitrag des baroklinen Vektors im Solenoidterm (letzter Term) kann in einer quasi-barotropen Approximation vernachlässigt werden, wird aber beispielsweise bei der Entstehung eines Tornados bedeutsam. Der erste Term auf der rechten Seite beschreibt das Dehnen (Stauchen) von vorhandenen Wirbelröhren durch die Divergenz (Konvergenz) des horizontalen Windfelds als Ausdruck der Drehimpulserhaltung (*Vortex Tube Stretching*). Der zweite Term beschreibt das Kippen von Wirbelröhren durch die vertikale Scherung des horizontalen Windfelds, da

$$\boldsymbol{\omega}_h \cdot \nabla w = -\mathbf{e}_z \cdot (\nabla w \times \mathbf{S}_h) \tag{2.51}$$

gilt (*Vortex Tilting*). Die Aufteilung des Windfelds in den Grundzustand eines rein höhenabhängigen horizontalen Windfelds und entsprechende dreidimensionale Störungen $\mathbf{v} = \overline{\mathbf{u}}(z) + \mathbf{v}'(\mathbf{x},t)$ impliziert, dass der Grundzustand keine vertikale Vorticity besitzt, sodass $\zeta = \zeta'(\mathbf{x},t)$ gilt. Das Einsetzen in Gleichung (2.50) und die Vernachlässigung aller nicht-linearen Störungsterme (Linearisierung) eliminiert den *Stretching*-Term, der unter den getroffenen Annahmen ein rein nicht-linearer Effekt ist und aus einem Ruhezustand mit $\zeta' = 0$ heraus ohnehin kaum Einfluss hat:

$$\frac{\overline{\mathrm{D}}\zeta'}{\mathrm{D}t} = \overline{\boldsymbol{\omega}}_h(z) \cdot \nabla w' = -\mathbf{e}_z \cdot \left[\nabla w' \times \overline{\mathbf{S}}_h(z)\right] \ . \tag{2.52}$$

Unter der Annahme, dass sich die Zelle mit einer konstanten Geschwindigkeit $\mathbf{c}_Z = c_{Z,x}\mathbf{e}_x + c_{Z,y}\mathbf{e}_y = const.$ verlagert, lautet Gleichung (2.52) in einem mit der Zelle bewegten natürlichen Koordinatensystem

$$\frac{\partial \zeta'}{\partial t} = -\left[\overline{\mathbf{u}}(z) - \mathbf{c}_Z\right]_s \frac{\partial \zeta'}{\partial s} - \left[\overline{\mathbf{u}}(z) - \mathbf{c}_Z\right]_n \frac{\partial \zeta'}{\partial n} + \overline{\omega}_s \frac{\partial w'}{\partial s} + \overline{\omega}_n \frac{\partial w'}{\partial n} \ . \tag{2.53}$$

Wird der tangentiale Einheitsvektor $\mathbf{e}_s$ parallel zum sturm-relativen Vektor $\overline{\mathbf{u}}_{sr} = \overline{\mathbf{u}} - \mathbf{c}_Z$, also $\overline{\mathbf{u}}_{sr} = |\overline{\mathbf{u}} - \mathbf{c}_Z|\mathbf{e}_s$, gelegt, vereinfacht sich dies zu

$$\frac{\partial \zeta'}{\partial t} = -|\overline{\mathbf{u}}_{sr}| \frac{\partial \zeta'}{\partial s} + \overline{\omega}_s \frac{\partial w'}{\partial s} + \overline{\omega}_n \frac{\partial w'}{\partial n} \ . \tag{2.54}$$

Darin nennt man den tangentialen Anteil der relativen Vorticity $\overline{\omega}_s = \overline{\boldsymbol{\omega}}_h \cdot \mathbf{e}_s$ *Streamwise Vorticity*, während der normale Anteil $\overline{\omega}_n = \overline{\boldsymbol{\omega}}_h \cdot \mathbf{e}_n$ als *Crosswise Vorticity* bezeichnet wird. Der normale Einheitsvektor $\mathbf{e}_n$ zeigt dabei orthogonal nach rechts vom tangentialen Einheitsvektor.

Das Wirbelkippen induziert in einem sich entwickelnden Aufwindbereich ($w' > 0$) unabhängig vom Verlauf der sturm-relativen Stromlinien Rotationspole (Vorticitymaxima) in dessen Randbereich. Exemplarisch seien Stromlinien in Richtung von $\overline{\mathbf{S}}_h$ betrachtet, sodass die horizontale Vorticity

der Umgebung $\overline{\boldsymbol{\omega}}_h = \overline{\omega}_n \mathbf{e}_n$ mit $\overline{\omega}_n < 0$ ist (reine *Crosswise Vorticity*), da der normale Einheitsvektor antiparallel zu den Wirbelröhren der Umgebung ist. Entlang der Stromlinien ist linksseitig (rechtsseitig) des Aufwindzentrums $\partial w'/\partial n > 0$ ($\partial w'/\partial n < 0$) mit dem größten Gradienten im Bereich der Wirbelröhre, die durch das Aufwindzentrum verläuft. Der Bereich links (rechts) des Aufwindzentrums erfährt daher wegen Gleichung (2.54) eine antizyklonale (zyklonale) Drehung, da $\partial\zeta'/\partial t > 0$ ($\partial\zeta'/\partial t < 0$) ist. Die Dipolachse, welche die Rotationspole verbindet, steht in der jeweiligen Höhenschicht senkrecht auf $\overline{\mathbf{S}}_h$. Ausgedrückt über das mit der vertikalen Windscherung in der Umgebung in Verbindung stehende Vorticityfeld $\overline{\boldsymbol{\omega}}_h$ ist die Dipolachse parallel zu den Wirbelröhren der Umgebung. Zum gleichen Ergebnis führt eine Betrachtung von Stromlinien, die von rechts senkrecht zum Scherungsvektor verlaufen ($\overline{\boldsymbol{\omega}}_h = \overline{\omega}_s \mathbf{e}_s$ mit $\overline{\omega}_s > 0$, reine *Streamwise Vorticity*).
Sobald vertikale Vorticity ζ' durch das Wirbelkippen generiert wird, wird sie sturm-relativ advehiert. Im Gegensatz zum Wirbelkippen ist der Verlauf der sturm-relativen Stromlinien entscheidend für den Effekt der Advektion. Im Fall reiner *Crosswise Vorticity* kommt es gemäß des Advektionsterms in Gleichung (2.54) zu einer Verschiebung der einzelnen Rotationspole entlang der Stromlinien stromabwärts von $\overline{\mathbf{S}}_h$. Der Grund hierfür ist, dass sich $-\partial\zeta'/\partial s$ entlang der Stromlinien linksseitig (rechtsseitig) des Aufwindzentrums verringert (vergrößert). Im Fall reiner *Streamwise Vorticity* kommt es zu einer Verschiebung des Rotationsdipols entlang der Stromlinie und damit senkrecht bezüglich der Richtung des Scherungsvektors. Daher vergrößert (verkleinert) sich $-\partial\zeta'/\partial s$ entlang der Stromlinie zwischen den Rotationspolen im Aufwindbereich (außerhalb des Aufwindbereichs). Dies führt folglich dazu, dass ein Rotationspol mit dem Aufwind in Phase gerät und sich eine Mesozyklone entwickeln kann, während im Fall reiner *Crosswise Vorticity* die Rotationspole im Randbereich des Aufwinds verbleiben. Aufgrund des größeren Betrags des Vorticitygradienten ist die Verschiebung der Rotationspole durch *Streamwise Vorticity* ausgeprägter als die Verschiebung durch *Crosswise Vorticity*.

Dynamische Modifikation des Aufwindbereichs und Zellaufteilung

Mit der Entwicklung eines Aufwindbereichs in einer vertikal gescherten Umgebung und den induzierten Rotationsdipolen gehen dynamische Druckstörungen p'_{dyn} einher, die – wie in Gleichung (2.35) dargestellt – vertikale Beschleunigungen hervorrufen können. Diese Druckstörungen setzen sich allgemein aus einem linearen und einem nicht-linearen Anteil zusammen:

$$p'_{dyn,l} \sim \overline{\mathbf{S}}_h(z) \cdot \nabla w' \,, \qquad p'_{dyn,nl} \sim -\zeta'^2 \,. \tag{2.55}$$

Je nachdem, wie die vertikale Windscherung ausgeprägt ist (wie groß also der Anteil von *Streamwise* und *Crosswise Vorticity* ist), haben diese Druckstörungen unterschiedliche Auswirkungen auf die weitere Entwicklung einer Superzelle (Weisman und Rotunno, 2000). Im Folgenden seien sie daher einmal für den Fall einer reinen Geschwindigkeitsscherung (gerader Hodograph) und einmal für den Fall einer Kombination aus Geschwindigkeits- und Richtungsscherung (gekrümmter Hodograph) diskutiert.

Gerader Hodograph

Der Fall eines geraden Hodographen als Folge reiner Geschwindigkeitsscherung wird hier o. B. d. A. durch eine reine Weströmung mit $\overline{\mathbf{u}}(z) = \overline{u}(z)\mathbf{e}_x$ betrachtet. Sowohl $\overline{\boldsymbol{\omega}}_h(z) = \mathbf{e}_y \partial\overline{u}/\partial z$ als auch $\overline{\mathbf{S}}_h(z) = \mathbf{e}_x \partial\overline{u}/\partial z$ ändern somit lediglich ihren Betrag mit der Höhe, nicht aber ihre Richtung. Eine sich entwickelnde Zelle bewegt sich parallel zum Scherungsvektor und somit in allen Höhenschichten senkrecht zu den Wirbelröhren der Umgebung.

Das Fehlen von Richtungsscherung bewirkt, dass sich durch das Wirbelkippen südlich (nördlich) des Aufwindbereichs ein hochreichender Bereich zyklonaler (antizyklonaler) Rotation ausbildet (Abbildung 2.6a). Aufgrund des

Fehlens von *Streamwise Vorticity* findet jedoch keine Verlagerung der Rotationspole senkrecht zum Scherungsvektor statt. Für die linearen Druckstörungen gilt:

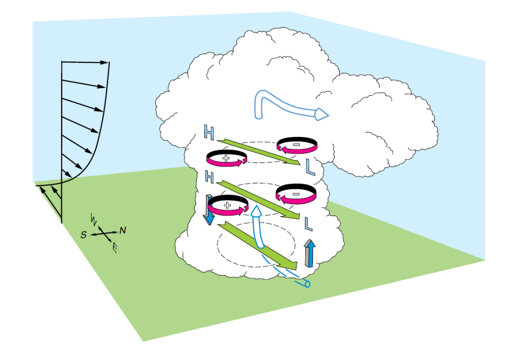
$$p'_{dyn,l} \sim \frac{\partial \overline{u}}{\partial z} \frac{\partial w'}{\partial x} \ . \qquad (2.56)$$

Somit wird ein Druckanstieg (Druckfall) an der Rückseite (Vorderseite) der Zelle hervorgerufen, welcher in mittleren Höhenschichten aufgrund der dort vorliegenden meist stärkeren Geschwindigkeitsscherung stärker ausfällt als in der unteren Troposphäre. Der resultierende Druckgradient ist demnach parallel zum Scherungsvektor $\overline{\mathbf{S}}_h$. Mit dem Druckfall auf der Vorderseite geht daher wegen Gleichung (2.35) eine positive Vertikalbeschleunigung einher, sodass es dort zu einem verstärkten Aufstieg einströmender Luftmassen kommt, der zu einer Neigung des Aufwindbereichs stromab des Scherungsvektors führt und die Verlagerung der Zelle unterstützt.
Die nicht-linearen Druckstörungen bewirken wegen der quadratischen Abhängigkeit von ζ' in den beiden den Aufwind flankierenden Rotationsbereichen einen Druckfall und daher eine positive Vertikalbeschleunigung, die aufgrund der meist stärkeren Geschwindigkeitsscherung in mittleren Höhenschichten stärker ausgeprägt ist. Neben dem ursprünglichen Aufwindbereich bilden sich dadurch zwei neue Aufwindbereiche im Bereich der Rotationszentren (Abbildung 2.6b). Die einströmenden Luftmassen divergieren zunehmend zu den beiden neuen Aufwindbereichen und schneiden den ursprünglichen Aufwindbereich von der Zufuhr warm-feuchter Luftmassen ab. Schließlich kann sich dort im Verlauf ein Abwindbereich entwickeln, der das Aufteilen (Split) in zwei achsensymmetrische Zellen einleitet (bezogen auf die Achse, die durch die Verlagerung des ursprünglichen Aufwindbereichs gegeben ist; z. B. Klemp und Wilhelmson, 1978a). Diese beiden Zellen bewegen sich in Richtungen etwas weiter nach links bzw. rechts als der ursprüngliche Aufwindbereich (*Left-* und *Right-Mover*). Dadurch entsteht ein zunehmender Anteil von *Streamwise Vorticity* für beide Zellen,

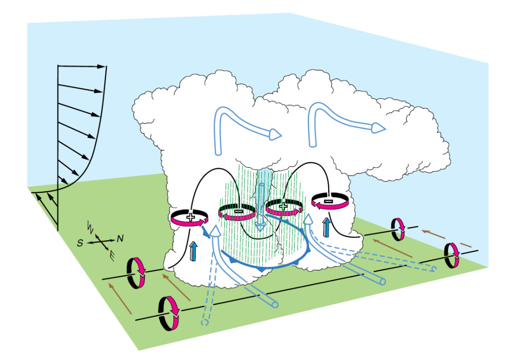
die in der Folge eine zyklonal (*Right-Mover*) bzw. antizyklonal (*Left-Mover*) rotierende Mesozyklone entwickeln können.

Gekrümmter Hodograph

Auf der Nordhemisphäre liegt aufgrund der synoptisch-skaligen Strömungsdynamik während konvektionsförderlicher Wetterlagen meist eine Rechtsdrehung des Windes mit der Höhe vor. In diesem Fall ist der Hodograph nach rechts gekrümmt und neben der Windscherung $\overline{\mathbf{S}}_h(z)$, die sich nun aus Geschwindigkeits- und Richtungsscherung zusammensetzt, ändern auch die horizontalen Wirbelröhren in der Umgebung $\overline{\boldsymbol{\omega}}_h(z)$ ihre Richtung mit der Höhe. Dadurch ändert sich die horizontale Achse der infolge des Wirbelkippens in Gleichung (2.54) entstehenden Rotationsdipole rechtsdrehend mit der Höhe. Die Achse steht jedoch immer senkrecht auf $\overline{\mathbf{S}}_h(z)$ in der jeweiligen Schicht (Abbildung 2.7).
Die linearen Druckstörungen bewirken bis in mittlere Troposphärenhöhen einen Druckanstieg (Druckfall) auf der rechten (linken) Flanke des Aufwindbereichs. In höheren Schichten ist es genau umgekehrt, sodass wegen Gleichung (2.35) rechtsseitig (linksseitig) eine positive (negative) Vertikalbeschleunigung hervorgerufen wird. Die nicht-linearen Druckstörungen bewirken bis in mittlere Troposphärenhöhen sowohl links- als auch rechtsseitig des Aufwindbereichs eine positive Vertikalbeschleunigung. Somit überlagern sich an der rechten Flanke lineare und nicht-lineare Druckstörungen konstruktiv, während sie sich an der linken Flanke entgegenwirken und weitgehend kompensieren. Daher entwickelt sich der rechtsseitige Aufwindbereich schneller, erreicht höhere Vertikalgeschwindigkeiten als der linksseitige und verlagert sich gleichzeitig in Richtung des ursprünglichen Aufwindzentrums. Dies liegt daran, dass der Verlagerungsvektor $\mathbf{c}_Z$ aufgrund der Richtungsscherung nicht parallel zum Scherungsvektor ist und die sturm-relativen Stromlinien besonders in den unteren Luftschichten eine Komponente parallel zu den Wirbelröhren der Umgebung haben, welche die sturm-relative Advektion von *Streamwise Vorticity* ermöglicht. Im mittleren und rechten



(a) Dynamik zu Beginn der Zellentwicklung bei geradem Hodographen



(b) Dynamik bei geradem Hodographen beim Split der Zelle

Abbildung 2.6: (a) Rotationsdipole (rot) mit vertikaler Vorticity ζ' ($+$ und $-$), lineare Druckstörungen $p'_{dyn,l}$ (H: positiv, L: negativ) und Druckgradienten (grüne Pfeile; parallel zu $\overline{\mathbf{S}}_h$) in einer sich entwickelnden Zelle bei reiner Geschwindigkeitsscherung (Vertikalprofil links). Blaue Pfeile kennzeichnen die dadurch hervorgerufenen Druckgradienten. Weiße Pfeile mit blauem Rand kennzeichnen mittlere Trajektorien von Luftteilchen. (b) Ähnliche Darstellung zu einem späteren Zeitpunkt, wenn sich induziert durch die nicht-linearen Druckstörungen zwei separate Aufwindbereiche gebildet haben. In der Mitte befindet sich ein Abwindbereich, der durch den fallenden Niederschlag initiiert wird. Zusätzlich sind die bodennahen Wirbelröhren durch ihre Rotationsachsen (schwarze Linien) und ihre Drehrichtung (rot) dargestellt. Das Kippen der Wirbelröhren ist anhand der hinteren Röhre erkennbar. Die Einströmrichtungen für die entstehenden *Left-* und *Right-Mover* sind mit gestrichelten Pfeilen angedeutet. Aus Markowski und Richardson (2010). © John Wiley & Sons (genehmigter Nachdruck).

Bereich der Zelle entsteht in der Folge eine zyklonal rotierende Mesozyklone, während eine weitere Abschwächung des linksseitigen Aufwindbereichs erfolgen kann. Kommt es infolge der nicht-linearen Druckstörungen zu einer Zellaufteilung, führen die linearen Druckstörungen demnach dazu, dass sich der *Right-Mover* erhält und sich die Mesozyklone sogar weiter verstärken kann, während sich der *Left-Mover* rasch abschwächt und dissipiert (z. B. Klemp und Wilhelmson, 1978a; Rotunno und Klemp, 1985; Houze et al., 1993; Markowski und Dotzek, 2011).
Dieser Effekt ist umso stärker, je ausgeprägter die rechtsdrehende Richtungsänderung des Windes mit der Höhe in der Umgebung ist. Numerische Simulationen zeigen, dass das Maß der Verstärkung des rechtsseitigen Aufwindbereichs nicht die Lebensdauer des *Right-Movers* nach der Zellteilung bestimmt (Weisman und Rotunno, 2000). Tatsächlich geben sie Hinweise, dass Aufwindbereiche auch nach einigen Stunden noch eine hohe Intensität aufweisen und damit eine lange Lebensdauer der Zelle bewirken können, wenn der Umgebungswind in der unteren Troposphäre um etwa 90° nach rechts dreht und in der mittleren Troposphäre näherungsweise richtungskonstant bleibt, wie sich auch durch Beobachtungen bestätigt (z. B. Burgess und Curran, 1985). Von einer großen Anzahl beobachteter Mesozyklonen auf der Nordhemisphäre (ohne Unterscheidung der Windprofile) rotieren insgesamt etwa 90 % aller Mesozyklonen zyklonal (Bunkers, 2002).

Erhaltung der Rotation des Aufwindbereichs

Entscheidend für eine lange Lebensdauer einer Superzelle ist die Erhaltung ihrer rotierenden Mesozyklone. Exemplarisch sei ein kritisches Höhenniveau betrachtet, in dem der sturm-relative Vektor $\overline{\mathbf{u}}_{sr}$ parallel zu den Wirbelröhren in der Umgebung ist, und das nach Lilly (1979) häufig im Reifestadium einer Superzelle in etwa 1,5 bis 2 km Höhe liegt. Unter der Annahme, dass die Zelle nach einiger Zeit einen Gleichgewichtszustand mit $\partial\zeta'/\partial t = 0$ erreicht hat,

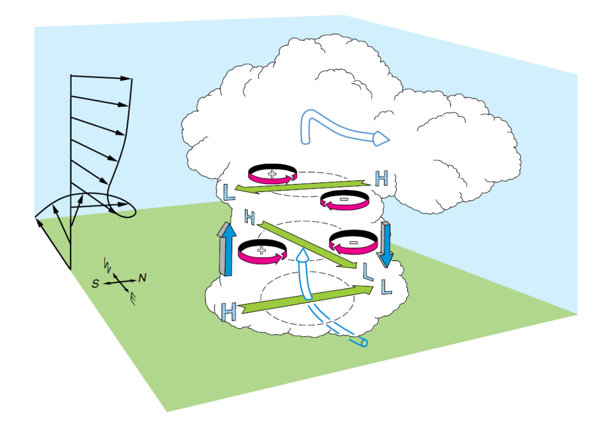


Abbildung 2.7: Wie Abbildung 2.6a, nur für den Fall eines gekrümmten Hodographen. Aus Markowski und Richardson (2010). © John Wiley & Sons (genehmigter Nachdruck).

gilt nach Gleichung (2.54) in diesem Höhenniveau

$$|\overline{\mathbf{u}}_{sr}|\frac{\partial \zeta'}{\partial s} = \overline{\omega}_s \frac{\partial w'}{\partial s} \ . \qquad (2.57)$$

Folgt man in diesem Niveau einer Stromlinie vom Bereich einströmender Luft in den Aufwindbereich, wird die Vorticity der Umgebung im Gleichgewichtszustand folglich komplett in vertikale Vorticity des Aufwinds mit derselben Rotationsrichtung umgewandelt (Davies-Jones, 1984). In Kombination mit dem nicht-linearen Effekt der Wirbelröhrendehnung, die gemäß Gleichung (2.50) mit steigender Rotationsstärke einen nicht zu vernachlässigenden Einfluss hat, wird insgesamt durch die sturm-relative Advektion der *Streamwise Vorticity* im Einströmbereich die Rotation des Aufwinds aufrechterhalten bzw. weiter verstärkt.

Sturm-relative Helizität (SRH)

Die SRH ist ein Maß für die sturm-relative *Streamwise Vorticity*, welches für wissenschaftliche Analysen sowie die Vorhersage von Gewittern verwendet wird (z. B. Droegemeier et al., 1993; Markowski und Richardson, 2010). Wegen des Einflusses der Verlagerungsrichtung der Zelle ist die SRH nicht Galilei-invariant und daher im mitbewegten Bezugssystem formuliert:

$$\begin{aligned} \mathrm{SRH}_{0-z'_\mathrm{o}} &= \int_{z'=0}^{z'_\mathrm{o}} (\overline{\mathbf{v}} - \mathbf{c}_Z) \cdot \overline{\boldsymbol{\varpi}} \, \mathrm{d}z' \\ &\overset{\overline{w}=0}{\approx} - \int_{z'=0}^{z'_\mathrm{o}} \mathbf{e}_z \cdot [(\overline{\mathbf{v}} - \mathbf{c}_Z) \times \mathbf{S}] \, \mathrm{d}z' \\ &= - \int_{z'=0}^{z'_\mathrm{o}} |\,(\overline{\mathbf{u}} - \mathbf{c}_Z) \times \mathbf{S}_h| \, \mathrm{d}z' \; . \end{aligned} \tag{2.58}$$

Darin bezeichnet z' die Höhe über Grund und z'_o den oberen Rand einer vertikalen Schicht. Der Verlagerungsvektor $\mathbf{c}_Z$ der Zellen kann mittels der Methode nach Bunkers et al. (2000) abgeschätzt werden:

$$\mathbf{c}_Z = \overline{\mathbf{u}}_{0-6\mathrm{km}} \pm D \left(\frac{\mathbf{e}_z \times \mathbf{S}_{0-6\mathrm{km}}}{|\mathbf{S}_{0-6\mathrm{km}}|} \right) \; . \tag{2.59}$$

Darin wurde D von den Autoren auf $7{,}5\,\mathrm{m\,s^{-1}}$ festgelegt, $\overline{\mathbf{u}}_{0-6\mathrm{km}}$ bezeichnet den vertikal gemittelten Horizontalwind in der Umgebung. Das Vorzeichen des zweiten Terms differenziert zwischen *Left-* und *Right-Movern*. Auch im Fall eines geraden Hodographen kann die SRH Werte ungleich Null erhalten, wenn die abgeschätzte Verlagerungsrichtung von der konstanten Windrichtung abweicht (wie z. B. bei den nach dem Aufteilen einer Zelle entstehenden *Left-* und *Right-Movern*). Große Werte der SRH sind ein Hinweis auf einen großen Anteil von *Streamwise Vorticity* im sturm-relativen Einströmbereich einer Zelle und können als Maß für die Wahrscheinlichkeit für einen zyklonal rotierenden Aufwindbereich interpretiert werden. Typische Werte, die beim tatsächlichen Auftreten von Superzellen berechnet werden, liegen bei

$SRH_{0-3km} \approx 250\,m^2\,s^{-2}$, während bei anderen Organisationsformen die Werte bei etwa $50\,m^2\,s^{-2}$ liegen (Thompson et al., 2003). Die SRH_{0-3km} ist darüber hinaus ein guter Indikator für die zu erwartende Länge der Zugbahn (und damit indirekt auch für die Lebensdauer) sowie die Hagelkorngröße einer Superzelle und zeigt sogar ein besseres diesbezügliches Unterscheidungsvermögen als die DLS (Kunz et al., 2020). Thompson et al. (2003) zeigten mit Hilfe von Daten aus Radiosondenaufstiegen in den USA, dass Superzellen dort bevorzugt in Umgebungen auftreten, in denen etwa $DLS > 20\,m\,s^{-1}$, $CAPE_{ML} > 1\,500\,J\,kg^{-1}$ und $BRN \approx 35$ ist. Aus Taszarek et al. (2020) lässt sich ableiten, dass die Trennwerte für die SRH_{0-3km}, DLS und $CAPE_{ML}$ in Europa teils deutlich niedriger liegen. Ausgeprägte, hochreichende Mesozyklonen sind hier jedoch auch weitaus seltener als in den USA.

2.2.4 Mesoskalige konvektive Systeme

Ein Mesoskaliges Konvektives System (MCS) ist ein Zusammenschluss vieler konvektiver Zellen zu einem konvektiven System, dessen Längenskala durch $L = 100\,km$ in mindestens eine horizontale Richtung gegeben ist (Trapp, 2013). Änderungen des horizontalen Strömungsfelds aufgrund der Coriolisbeschleunigung können auf dieser Längenskala mit einer charakteristischen Windgeschwindigkeit von $U = 10\,m\,s^{-1}$ nach Gleichung (2.27) dieselbe Größenordnung erreichen wie die advektiven Änderungen des Windfelds[6]:

$$\begin{aligned} \mathcal{O}\left([\mathbf{v}\cdot\nabla]\,\mathbf{u}\right) &= \mathcal{O}\left(f\times\mathbf{u}\right) \\ \Longrightarrow \frac{U^2}{L} &= |f|U\ . \end{aligned} \tag{2.60}$$

[6] Mittlere Vertikalbewegungen auf dieser Längenskala sind deutlich schwächer als die horizontale Strömung und resultieren daher in vernachlässigbaren Beiträgen zur horizontalen Coriolisbeschleunigung.

Darin steht $f = 2\Omega \sin(\phi)$ mit dem Betrag der Winkelgeschwindigkeit der Erdrotation $\Omega = 2\pi(24\,\mathrm{h})^{-1}$ und dem Breitengrad ϕ für den Coriolisparameter, der in mittleren Breiten bei $\phi = 45°$ etwa $|f| = 10^{-4}\,\mathrm{s}^{-1}$ beträgt. Die entsprechende Zeitskala ist dementsprechend $T = L\,U^{-1} = |f|^{-1} \approx 3\,\mathrm{h}$. MCS entwickeln häufig eine mesoskalige Zirkulation, die dazu führt, dass solche Systeme über Zeitskalen existieren können, die bis zu einer Größenordnung über dieser Zeitskala liegen (ca. 1 Tag).

Klassifikationen

MCS sind allgemein durch ein großes, zusammenhängendes Niederschlagsgebiet gekennzeichnet. Ein Teil ist dabei durch konvektive Niederschläge geprägt, der andere durch stratiforme Niederschläge (s. u.). Grundsätzlich werden zwei verschiedene Typen von MCS nach ihrem Entstehungsprozess unterschieden. Typ 1 MCS entstehen bereits kurz nach einer durch großräumige Hebung hervorgerufenen verbreiteten Auslösung konvektiver Zellen. Dies kann beispielsweise im Bereich des isentropen Aufgleitens über eine synoptisch-skalige Front oder im Bereich von bodennahen Konvergenzen geschehen, besonders wenn die Konvektionshemmung (CIN; vgl. Kapitel 2.1.2) nicht zu stark ausgeprägt ist (Markowski und Richardson, 2010). Typ 2 MCS bilden sich durch den Zusammenschluss des *Cold Pools* bereits existierender Einzel-, Multi- oder Superzellen (*Upscale Growth*). Sie entwickeln sich insbesondere an Tagen, an denen die Konvektionsauslösung an den Tagesgang der solaren Einstrahlung gekoppelt ist, folglich häufig in den Abendstunden. Der Fortbestand der MCS über die Nacht wird durch die Präsenz nächtlicher Grenzschichtstrahlströme unterstützt, die ein Windmaximum am Oberrand der Grenzschicht in $1-2\,\mathrm{km}$ Höhe zur Folge haben (Trapp, 2013). Deren vertikale Komponente trägt zu vertikalen Auslenkungen von Luftpaketen in Richtung des NFK bei. Gleichzeitig bedeutet ein solches Windmaximum eine vertikale Scherung des Horizontalwinds an den Rändern des Strahlstroms, welche die Organisation des konvektiven Systems unterstützt (s. u.). Darüber

hinaus erhöht ein solcher Strahlstrom die Massenkonvergenz im Bereich der Front des MCS, die für das weitere Bestehen des MCS förderlich ist. Bisweilen tragen auch orografisch oder selbst-induzierte Schwerewellen zur Auslenkung von Luftpaketen bei. Unter solch günstigen Umgebungsbedingungen können sich MCS die ganze Nacht hindurch am Leben erhalten.
Entwickelt ein MCS eine sehr große horizontale Ausdehnung, wird es auch als Mesoskaliger Konvektiver Komplex (MCC) bezeichnet. MCC stellen somit eine Unterklasse der MCS dar, die meist eine besonders lange Lebensdauer erreichen. Maddox (1980) legte folgende Klassifikationskriterien für einen MCC basierend auf abgeleiteten Beobachtungsgrößen aus Infrarot-Satellitenbildern fest: 1) Der Wolkenschirm muss auf einer Fläche von mehr als 100 000 km^2 Temperaturen niedriger als $\vartheta = -32\,°C$ vorweisen; 2) Zugleich muss der Wolkenschirm auf einer Fläche von mehr als 50 000 km^2 Temperaturen niedriger als $\vartheta = -52\,°C$ vorweisen; 3) Die Kriterien 1 und 2 müssen über eine Zeitspanne von mindestens 6 h erfüllt sein; 4) Für das Verhältnis der horizontalen Achsen des Wolkenschirms muss gelten $d_{kurz}\ d_{lang}^{-1} \geq 0{,}7$. Das System darf demnach nicht zu stark von der Kreisform abweichen[7]. Aufgrund der großen raum-zeitlichen Skala führen MCC häufig zu lang anhaltenden, mäßigen bis starken Regenfällen und können daher hohe Regensummen verursachen (z. B. Wilhelm et al., 2021). Insbesondere können MCC an ihrer Front starke konvektive Windböen hervorrufen. Unter der Wirkung der Corioliskraft kommt es im Zentrum des Systems häufig zu Druckfall, sodass ein mesoskaliger konvektiver Wirbel entsteht (Davis und Trier, 2007; vgl. Schmidberger, 2018).

[7] Maddox (1980) bezeichnet dieses Verhältnis als Exzentrizität. In der Mathematik wird die (numerische) Exzentrizität einer Ellipse mit den Achsen d_{lang} und d_{kurz} hingegen über $\varepsilon = (1 - d_{kurz}^2\ d_{lang}^{-2})^{0,5}$ definiert. Für einen Kreis gilt demnach $\varepsilon = 0$ und ein übertragenes Kriterium für MCC würde lauten: $\varepsilon < 0{,}714$.

Rotunno-Klemp-Weisman-Theorie für Gewitterlinien

Zu den Typ 1 MCS zählen Gewitterlinien (*Squall Lines*), die sich lediglich in eine horizontale Richtung über die charakteristische Längenskala erstrecken. In Abhängigkeit von den sturm-relativen Winden in der mittleren und oberen Troposphäre befinden sich die stratiformen Niederschläge einer Gewitterlinie eher auf der Rück- oder Vorderseite (*Trailing* bzw. *Leading Stratiform Precipitation*) bzw. entlang der langen horizontalen Achse (*Parallel Stratiform Precipitation*). Am häufigsten beobachtet wird die Kombination aus konvektivem Niederschlag an der Vorderseite und stratiformem Niederschlag auf der Rückseite der Linie (ca. $60 - 80\,\%$; Parker und Johnson, 2000). In Abhängigkeit von der bodennahen Luftfeuchte und weiteren Faktoren weisen Gewitterlinien entweder einen zusammenhängenden Aufwindbereich (*Slab-like Updraft*) oder mehrere zellartige Aufwindbereiche (*Cellular Updraft*) auf (Markowski und Richardson, 2010). Besonders Gewitterlinien mit einem zusammenhängenden Aufwindbereich können als zweidimensionales konzeptionelles Modell vereinfacht dargestellt werden. Die relevanten Prozesse können so in einem 2D-Vertikalschnitt entlang der kurzen horizontalen Achse der Gewitterlinie mit der Längenskala $L' \approx 10\,\text{km}$ betrachtet werden.

Die zeitliche Entwicklung einer Gewitterlinie kann in einem solchen zweidimensionalen Modell am besten mit Hilfe der Rotunno-Klemp-Weisman-Theorie verstanden werden (z. B. Rotunno et al., 1988; Weisman et al., 1988; Weisman und Rotunno, 2004). In der 2D-Geometrie werden in dieser Theorie neben Reibungseffekten auch Corioliseffekte vernachlässigt (da $L' < L$ ist). Durch die Anwendung der Rotation auf die Impulsgleichung (2.27) und der anschließenden Projektion auf die Vorticitykomponente senkrecht zum Vertikalschnitt lässt sich eine horizontale Vorticitygleichung herleiten. Unter der Annahme einer inkompressiblen Strömung führt die Integration dieser Vorticitygleichung über ein Kontrollvolumen im Bereich des niederschlagsinduzierten *Cold Pools* zu einem Gleichgewicht der Effekte der vertikalen

Windscherung in der Umgebung und des vertikal integrierten Auftriebs im Bereich des Kontrollvolumens. Die Windscherung kann dabei durch die Differenz des Winds zwischen der Ober- und Unterkante am stromab befindlichen Rand des Kontrollvolumens Δu dargestellt werden. Mit dem vertikal integrierten Auftrieb im Bereich des *Cold Pools* $\mathscr{B}$ folgt:

$$\Delta u = \sqrt{2|\mathscr{B}|} \, . \tag{2.61}$$

Dieses Gleichgewicht stellt ein Kriterium für eine optimale, sich selbst erhaltende Gewitterlinie dar und kann als Gleichgewicht zwischen der (positiven) horizontalen Vorticity, die durch die Windscherung in den unteren Schichten hervorgerufen wird, und der (negativen) horizontalen Vorticity, die mit dem negativen Auftrieb $\mathscr{B}$ einhergeht, interpretiert werden (Abbildung 2.8). Zu Beginn des Lebenszyklus einer Gewitterlinie, wenn sich die entwickelnden Zellen im Übergang vom Cumulus- zum Reifestadium befinden, beginnen die entstehenden Abwindbereiche einen *Cold Pool* zu bilden. Hierbei ist $\sqrt{2|\mathscr{B}|} \ll \Delta u$ und der Aufwindbereich ist durch die Dominanz der vertikalen Windscherung bzw. der positiven Vorticity stromab des Scherungsvektors geneigt (Abbildung 2.8a). Fällt Niederschlag aus, geschieht dies meist im Einströmbereich der warm-feuchten Luft, was im Fall einer ausgeprägten einströmenden Schicht dort zu einer (vorübergehenden) Destabilisierung führt (nicht gezeigt). Im Fall einer flachen einströmenden Schicht kann der Niederschlag – ähnlich wie bei dem Lebenszyklus einer isolierten Einzelzelle – den Aufwindbereich von den einströmenden Luftmassen abschneiden und so das Ende des Lebenszyklus einleiten.

Wird mit der Zeit jedoch ein Gleichgewicht mit $\sqrt{2|\mathscr{B}|} \approx \Delta u$ erreicht, balancieren sich die positiven und negativen Vorticitybeiträge und erhalten so die Gewitterlinie durch die wiederholte Bildung hochreichender Zellen an der Vorderseite des Systems aufrecht (Abbildung 2.8b). Bryan et al. (2006) zeigten in einer umfangreichen Modellierungsstudie, dass die intensivsten

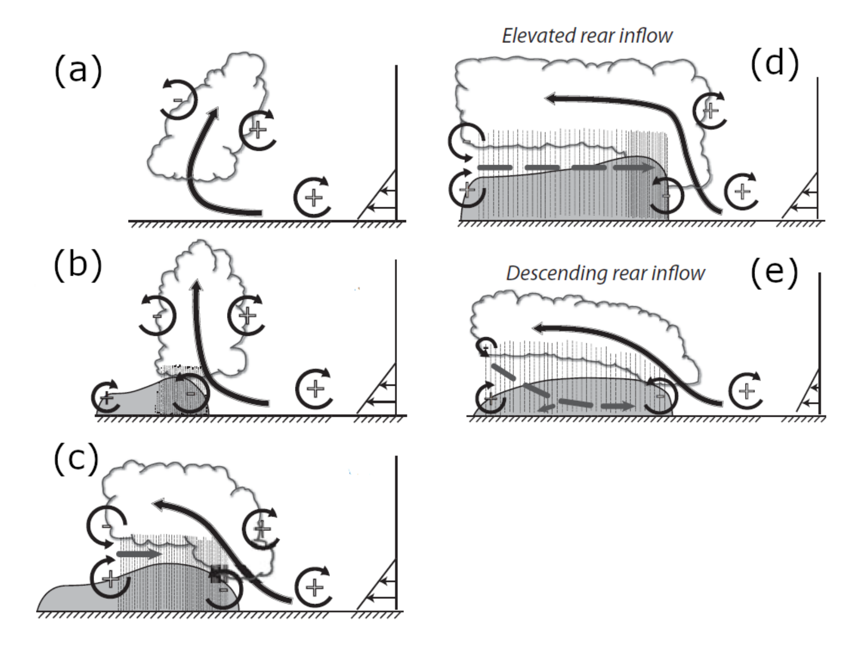


Abbildung 2.8: Vertikalschnitt durch eine Gewitterlinie zur Erläuterung der Rotunno-Klemp-Weisman-Theorie nach Weisman (1992) und Weisman und Rotunno (2004), übernommen von Trapp (2013) und leicht modifiziert. Die Vorticity-Balance ist durch rotierende Pfeile mit dem jeweiligen Vorzeichen der Vorticity illustriert. Der schwarze Pfeil kennzeichnet Trajektorien von einströmenden Luftteilchen. Der *Cold Pool* ist durch die graue Fläche gekennzeichnet. Zurückströmende Luftmassen im Bereich des *Cold Pools* sind durch graue Pfeile angedeutet. (a) $\sqrt{2|\mathscr{B}|} \ll \Delta u$; (b) $\sqrt{2|\mathscr{B}|} \approx \Delta u$; (c) $\sqrt{2|\mathscr{B}|} \gg \Delta u$; (d) und (e): Erweiterung der Theorie um den *Rear Inflow Jet* für (d) den Fall eines starken *Rear Inflow Jets* mit hochreichendem *Cold Pool* und (e) den Fall eines schwachen *Rear Inflow Jets* mit einem flachen *Cold Pool*. © Cambridge University Press (genehmigter Nachdruck).

Gewitterlinien mit den stärksten bodennahen Windböen und höchsten Niederschlagssummen im Bereich dieses Gleichgewichts auftreten. Für Gewitterlinien, die lange Zeit dieses Gleichgewicht aufrechterhalten, ist der Zusammenhang mit einer insgesamt längeren Lebensdauer des Systems naheliegend, es bedarf jedoch noch weiterer Untersuchungen. Im weiteren Verlauf überwiegt aufgrund der Verstärkung des *Cold Pools* negative Vorticity. Einströmende Luftmassen werden automatisch stromaufwärts des Scherungsvektors transportiert, wodurch es zu einer weiteren Verstärkung des

Vorticity-Ungleichgewichts kommt mit $\sqrt{2|\mathscr{B}|} \gg \Delta u$ (Abbildung 2.8c). Dieser Mechanismus erklärt somit auch die Entstehung der am häufigsten beobachteten Niederschlagsverteilung (*Trailing Stratiform*; s. o.). Wird der Bereich negativer Vorticity zu dominant, beginnt das Dissipationsstadium der Gewitterlinie.
Eine Erweiterung der vorgestellten Theorie, die insbesondere im Reifestadium der Gewitterlinien relevant ist, berücksichtigt zusätzlich eine Vorticity-Balance auf der Rückseite der Gewitterlinie (Abbildung 2.8c−e). Bodennah wird durch den *Cold Pool* dort positive Vorticity generiert, während durch die stromaufwärts einströmenden Luftmassen oberhalb des *Cold Pools* negative Vorticity erzeugt wird. Diese Kombination unterstützt die Entwicklung eines von der Rückseite einströmenden Strahlstroms (*Rear Inflow Jet*). Dieser verstärkt an der Linienfront je nach Ausprägung die Konvergenz und das Aufsteigen der einströmenden Luft und kann damit zum Erhalt des konvektiven Systems beitragen, sodass sich ein neues Vorticitygleichgewicht einstellt. Die negative Vorticity des *Cold Pools* wird dann durch die positive Vorticity aus der Windscherung der Umgebung und dem *Rear Inflow Jet* balanciert.

Die Verlagerungsgeschwindigkeit einer Gewitterlinie (bzw. allgemein eines MCS) setzt sich aus der mittleren Strömung und der Entwicklungsrichtung der sich neu bildenden Zellen an der Front des MCS zusammen – wie für kleinskaligere Multizellen in Kapitel 2.2.2 beschrieben. Ist der *Rear Inflow Jet* sehr stark ausgeprägt, kann er Geschwindigkeiten erreichen, welche die Verlagerungsgeschwindigkeit der Gewitterlinie überschreiten. Dies hat zur Folge, dass der *Rear Inflow Jet* die Front bogenförmig stromabwärts (konkav) deformiert. Im Niederschlagsradar ist diese Form besonders gut zu erkennen und wird daher auch als Bogenecho (*Bow Echo*) bezeichnet (Fujita, 1978; Weisman, 1993). Es kommt vor, dass mehrere solcher Bogenechos entlang der Linienfront als wellenartige Struktur im Radarbild zu erkennen sind (*Line Echo Wave Pattern*; Nolen, 1959). Durch die Zunahme

der Vorticity an den Rändern des Bogens kann sich das Bogenecho im weiteren Verlauf zu einer Art Komma deformieren und dort, wo zyklonale Vorticity generiert wird, ein Hakenecho bilden. Durch den starken *Rear Inflow Jet* sind besonders bei solchen Entwicklungen extrem hohe Windgeschwindigkeiten an der Böenfront sowie starke Fallböen möglich (z. B. Hamilton, 1970; Mathias et al., 2017).

2.3 Atmosphärische Umgebungsvariablen, Kenngrößen und konvektive Indizes

Wie in den Kapiteln 1 und 2.1 beschrieben sind ein ausreichendes Feuchteangebot in der unteren Troposphäre, eine labile Schichtung der Atmosphäre und ein Mechanismus, der vertikal ausgelenkten Luftpaketen einen freien Aufstieg durch thermischen Auftrieb ermöglicht, notwendige Voraussetzungen für die Entstehung hochreichender Feuchtkonvektion. Die ersten beiden Voraussetzungen lassen sich dabei durch verschiedene Kenngrößen und sogenannte konvektive Indizes quantifizieren (vgl. Huntrieser et al., 1997; Haklander und van Delden, 2003; Kunz, 2007). Der vorherrschende Hebungsmechanismus hingegen lässt sich nicht über solche Indizes ausdrücken. Die für die Organisationsform konvektiver Zellen relevante vertikale Windscherung der Umgebung lässt sich wiederum über dynamische Kenngrößen wie beispielsweise die DLS oder die SRH darstellen.
In der vorliegenden Arbeit finden neben den bereits eingeführten Größen CAPE, CIN, BRN, DLS und SRH und grundlegenden Variablen wie z. B. Lufttemperatur, Windgeschwindigkeit und Luftfeuchte auch weitere Kenngrößen und Indizes Verwendung. Meist wird in den folgenden Kapiteln der Einfachheit halber generell von Umgebungsvariablen gesprochen, womit alle

atmosphärischen Variablen, Kenngrößen und Indizes gemeint sind[8]. Wichtige Umgebungsvariablen, die in den vorherigen Kapiteln noch nicht eingeführt wurden, werden im Folgenden beschrieben. Weitere Indizes sind in Anhang A dargestellt. Eine große Zahl von Studien aus vielen Teilen der Erde konnte einen statistischen Zusammenhang zwischen solchen Kenngrößen und dem Auftreten konvektiver Zellen und/oder bestimmter Begleiterscheinungen wie Tornados, Blitzen, Hagel, Starkregen oder Sturmböen herstellen (s. u.).

Mittlerer vertikaler Temperaturgradient (*Lapse Rate*)

Der Terminus *Lapse Rate* wird meist als Bezeichnung für den vertikalen Temperaturgradienten verwendet, wobei die *Lapse Rate* bei einer Temperaturabnahme mit der Höhe ein positives Vorzeichen erhält (und damit γ aus Kapitel 2.1.2 entspricht; z. B. Markowski und Richardson, 2010). Zudem dient der Begriff häufig dazu, einen mittleren vertikalen Temperaturgradienten über eine bestimmte vertikale Schicht zu charakterisieren, sodass die *Lapse Rate* ein Maß für die thermische Stabilität innerhalb dieser Schicht ist. Zwischen einem bestimmten Höhenniveau z_m und dem Grund bestimmt sie sich über

$$\mathrm{LR}_{0-z_m} = \frac{T_0 - T_m}{z_m} \, , \tag{2.62}$$

wobei T_0 die bodennahe Temperatur und T_m die Temperatur im Höhenniveau z_m kennzeichnet. Zur Bestimmung der *Lapse Rate* in der mittleren Troposphäre ist es gebräuchlich, die Schicht durch zwei Druckniveaus p_m und p_n

[8] Den Erläuterungen zur Theorie eines gehobenen Luftpakets in Kapitel 2.1.2 entsprechend ist der Wert einiger Variablen wie z. B. der CAPE und der CIN abhängig von den Startbedingungen und dem Ausgangsniveau des Luftpakets, welches zur Berechnung angenommen wird. Diese werden in den nachfolgenden Kapiteln durch einen tiefgestellten Index an die Akronyme angehängt, z. B. $\mathrm{CAPE}_{\mathrm{MU}}$ für die CAPE basierend auf einem Luftpaket, das in der Schicht mit der höchsten pseudopotentiellen Temperatur startet. Weitere Variablen erhalten einen Index, welcher die jeweilige vertikale Höhen- oder Druckschicht kennzeichnet, z. B. $T_{850\mathrm{hPa}}$ für die 850 hPa Temperatur oder $\overline{\mathbf{U}}_{0-6\mathrm{km}}$ für den zwischen 0 und 6 km gemittelten Horizontalwind.

abzugrenzen ($p_m > p_n$). Dann muss zunächst jeweils das entsprechende Höhenniveau z_{p_m} bzw. z_{p_n} bestimmt werden. Die *Lapse Rate* ergibt sich in diesem Fall über

$$\mathrm{LR}_{p_m - p_n} = -\frac{T_{p_m} - T_{p_n}}{z_{p_m} - z_{p_n}} , \tag{2.63}$$

mit den entsprechenden Temperaturwerten an den Schichtgrenzen T_{p_m} und T_{p_n}.

Lifted Index (LI)

Der LI als Maß für latente Instabilität bestimmt sich über die Differenz zwischen der Umgebungstemperatur im 500 hPa Niveau und der Temperatur eines fiktiven aufsteigenden Luftpakets in diesem Niveau:

$$\mathrm{LI} = T_{U,500\mathrm{hPa}} - T_{P,500\mathrm{hPa}} . \tag{2.64}$$

Wird ein Luftpaket angenommen, dessen Startbedingungen durch die bodennahen Werte vorgegeben werden, so bezeichnet man den LI nach Galway (1956) als *Surface Lifted Index* (SLI). Dieser sowie zwei weitere LI, die auf einem ML-Luftpaket (50 und 100 hPa Schichtdicke) basieren, finden in der vorliegenden Arbeit Verwendung. Ist $\mathrm{LI} < 0$, so herrscht in der mittleren Troposphäre ein positiver Auftrieb vor. Der LI wurde in vielen Studien zur Charakterisierung latenter Instabilität verwendet und gilt als guter Indikator für das Auftreten konvektiver Zellen (z. B. Haklander und van Delden, 2003; Kunz, 2007).

Vertikal integrierter Wasserdampfgehalt (IWV)

Der vertikal integrierte Wasserdampfgehalt ist ein Maß für die in einer Einheits-Luftsäule enthaltene Menge an Wasserdampf. Er lässt sich über

$$\mathrm{IWV} = \int_{z_B}^{z_o} \rho_v \mathrm{d}z \overset{(2.11)}{=} \int_{z_B}^{z_o} \frac{p}{R_d T} \frac{\frac{R_d}{R_v} r_v}{r_v + \frac{R_d}{R_v}} \mathrm{d}z \tag{2.65}$$

mit der Höhe des Erdbodens z_B und einer oberen Grenze z_o berechnen, welche je nach Definition durch den Oberrand der Troposphäre bzw. der gesamten Atmosphäre gegeben ist. Die Beiträge aus den Sphären oberhalb der Tropopause sind dabei jedoch vernachlässigbar klein. Der IWV kann beispielsweise anhand von Satellitenbeobachtungen sowie in Atmosphärenmodellen bestimmt werden. Bei letzteren erfolgt eine Diskretisierung des Integrals in eine Summe über die Modellschichten bezogen auf die jeweilige Vertikalkoordinate. Je höher der IWV ist, desto mehr Wasserdampf steht potentiell zur Kondensation zur Verfügung, sollte ein aufsteigendes Luftpaket das HKN bzw. KKN erreichen. Typische Werte in Mitteleuropa während konvektiv geprägter Wetterlagen sind $\mathrm{IWV} = 25 - 35\,\mathrm{kg\,m^{-2}}$. Der IWV weist beispielsweise einen Zusammenhang zur Häufigkeit von großem Hagel auf (z. B. Cao, 2008). Zudem ist er eng mit dem vertikal integrierten Flüssigwassergehalt verbunden, welcher über den Radarreflektivitätsfaktor Z aus 3D-Radardaten abgeschätzt werden kann (s. Kapitel 4.1.1; Greene und Clark, 1972). In der Praxis kann dieser gut zwischen schweren und weniger schweren konvektiven Zellen unterscheiden (Kitzmiller et al., 1995).

Supercell Composite Parameter (SCP)

Der SCP betrachtet zwei wichtige Faktoren für die Entstehung bzw. Organisation konvektiver Zellen in Kombination: Instabilität und vertikale Windscherung (Thompson et al., 2003). Die in der vorliegenden Arbeit verwendete Version orientiert sich an der Definition von Gensini und Tippett (2019):

$$\mathrm{SCP} = \frac{\mathrm{CAPE_{MU}}}{1\,000\,\mathrm{J\,kg^{-1}}} \frac{\mathrm{SRH_{0-3km}}}{100\,\mathrm{m^2\,s^{-2}}} \frac{\mathrm{DLS}}{20\,\mathrm{m\,s^{-1}}}\ . \tag{2.66}$$

Der SCP als multiplikatives Maß erreicht den Wert 1, wenn die CAPE_{MU}, $\text{SRH}_{0-3\text{km}}$ und DLS typische Schwellenwerte für Umgebungsbedingungen annehmen, die für die Bildung von Superzellen förderlich sind (vgl. Kapitel 2.2.3). In wissenschaftlichen Studien werden Umgebungsbedingungen konvektiver Zellen sehr häufig in einer kombinierten Betrachtungsweise von vertikaler Windscherung und Instabilität untersucht sowie einige weitere kombinierte Indizes vorgeschlagen und angewendet (z. B. Brooks et al., 2007; Groenemeijer und van Delden, 2007; Púčik et al., 2015; Sherburn et al., 2016; Westermayer et al., 2017). Manche kombinierte Indizes wie der SCP sind darüber hinaus in einigen operationellen Vorhersagesystemen von nationalen Wetterdiensten implementiert.

Significant Hail Parameter (SHIP)

Ein weiterer kombinierter Index ist der SHIP, der neben drei unterschiedlichen Maßen der Instabilität und der vertikalen Windscherung zusätzlich noch die Feuchte in Form des Wasserdampfmischungsverhältnisses am HKN berücksichtigt[9]:

$$\text{SHIP} = \frac{1}{42\,000\,000} \frac{\text{CAPE}_{\text{MU}}}{\text{J}\,\text{kg}^{-1}} \frac{\text{DLS}}{\text{m}\,\text{s}^{-1}} \frac{r_{v,\text{HKN}}}{\text{g}\,\text{kg}^{-1}} \cdot \frac{\text{LR}_{700-500\text{hPa}}}{\text{K}\,\text{km}^{-1}} \frac{273{,}16\,\text{K} - T_{500\text{hPa}}}{\text{K}} \,. \tag{2.67}$$

Je größer die Werte des SHIP, desto förderlicher sind die Umgebungsbedingungen für (großen) Hagel. Der Normierungsfaktor kann in ähnlicher Weise wie für den SCP verstanden werden, sodass SHIP = 1 einen Trennwert zwischen Umgebungsbedingungen darstellt, bei denen vornehmlich kleinerer oder größerer Hagel beobachtet wird. Neben der operationellen

[9] `https://www.spc.noaa.gov/exper/mesoanalysis/help/help_sigh.html`

Wettervorhersage erweist sich auch der SHIP in wissenschaftlichen Studien zu Hagelumgebungen als sehr nützlich (z. B. Prein und Holland, 2018; Czernecki et al., 2019; Tang et al., 2019).

2.4 Lebenszyklen konvektiver Zellen und Multi-Daten-Ansatz

Im Anschluss an die Entwicklung konzeptioneller Modelle von konvektiven Zellen auf der Basis von Beobachtungsdaten widmeten sich in der zweiten Hälfte des 20. Jahrhunderts viele wissenschaftliche Studien dank der fortschreitenden Entwicklung der Computertechnik und der numerischen Modellierung vermehrt der numerischen Simulation der zeitlichen Entwicklung konvektiver Zellen, um die theoretischen Konzepte zu überprüfen und ein tiefergehendes Prozessverständnis zu generieren (vgl. Kapitel 2.2; z. B. Orville und Sloan, 1970; Wilhelmson, 1974; Klemp und Wilhelmson, 1978b; Fovell und Tan, 1998). Heutzutage beschäftigen sich viele Studien – wie auch die vorliegende Arbeit – darüber hinaus mit dem *Nowcasting* konvektiver Zellen, welches unter anderem für den automatisierten Warnprozess der Wetterdienste von großer Bedeutung ist (s. u.; vgl. Kapitel 1). Algorithmen, die konvektive Zellen automatisch in Produkten aus Fernerkundungsmethoden (Satelliten-, Radar- und/oder Blitzdaten) detektieren und verfolgen, spielen dabei eine wichtige Rolle (Zellverfolgungsalgorithmen). Basierend auf den Fernerkundungsdaten und den daraus abgeleiteten Daten der Zellverfolgungsalgorithmen sind nicht nur nachträgliche Untersuchungen der Lebenszyklen einzelner konvektiver Ereignisse möglich (z. B. Höller et al., 1994; Schmidt et al., 2012; Wapler et al., 2015; Kunz et al., 2018), sondern auch statistische Untersuchungen der Eigenschaften einer großen Anzahl konvektiver Zellen (z. B. Davini et al., 2012; Meyer et al., 2013; Wapler, 2017;

Schmidberger, 2018; Zöbisch, 2020; Wapler, 2021). Damit können detaillierte Charakterisierungen der Lebenszyklen konvektiver Zellen mit Hilfe unterschiedlicher Beobachtungsgrößen vorgenommen werden, welche wiederum zur Verbesserung von *Nowcasting*-Verfahren genutzt werden können. Viele *Nowcasting*-Verfahren und insbesondere Zellverfolgungsalgorithmen behandeln detektierte Zellen in der Regel gleich, unabhängig von der Organisationsform oder dem genauen Entwicklungsstadium (Lebenszyklusphase) der konvektiven Systeme. Hier dienen die jeweiligen Beobachtungsgrößen als Grundlage für die automatische Zellanalyse und die Abschätzung der weiteren Entwicklung (*Nowcasts*).

In einer aktuellen Studie weisen Zöbisch et al. (2020) darauf hin, dass die Abschätzung der Lebensdauer konvektiver Zellen (bzw. der Zeit vom jeweiligen Detektionszeitpunkt bis zur finalen Dissipation der Zellen [verbleibende Lebensdauer]) unabhängig von der Organisationsform eine zentrale Herausforderung für *Nowcasting*-Verfahren der aktuellen Generation darstellt. Sie geben darüber hinaus einen umfangreichen, wenn auch nicht allumfassenden Überblick über eine Reihe von Studien, die sich mit der Untersuchung der Lebenszyklen konvektiver Zellen basierend auf Fernerkundungs- und Modelldaten beschäftigen. Die Berichte der zurückliegenden *Nowcasting*-Konferenzen geben zudem einen Überblick über die aktuellen Schwerpunkte im Bereich des *Nowcastings*, welche die Wetterdienste als relevant und entscheidend für die potentielle Verbesserung von *Nowcasting*-Verfahren identifiziert haben (z. B. Wapler, 2017; Schmid et al., 2019). In Anlehnung an die genannten Veröffentlichungen seien einige wichtige Aspekte zu den Lebenszyklen konvektiver Zellen in Fernerkundungsdaten kurz dargelegt, ohne auf die Funktionsweise der Fernerkundungsmethoden und jeden einzelnen Aspekt sowie physikalische Interpretationen der jeweiligen Zusammenhänge im Detail einzugehen. Vielmehr stellen die folgenden Abschnitte eine kurze Synthese dieser Aspekte dar, die eine Brücke zwischen den theoretischen Grundlagen der

vorangegangenen Kapitel und den Methoden und Analysen der folgenden Kapitel schlagen.

Mit Hilfe von Satelliten können konvektive Zellen bereits früh in ihrem Cumulusstadium als schnell anwachsende Wolke identifiziert werden, deren optische Transparenz mit der Zeit abnimmt und deren Oberrand eine rasche Temperaturabnahme verzeichnet (z. B. Mecikalski et al., 2011; Senf und Deneke, 2017; Zöbisch et al., 2020). Senf et al. (2015) zeigen für einige Fallbeispiele in Mitteleuropa, dass die Temperaturabnahme dabei keinen systematischen Zusammenhang zur vorhandenen latenten Instabilität vermuten lässt und schreiben dies dem komplexen Terrain Mitteleuropas und den damit verbundenen Auslösemechanismen von Konvektion zu. Wie auch von Mecikalski et al. (2013) dargelegt, wird aufgrund der besseren Beobachtungslage durch die Verfügbarkeit hochaufgelöster Satellitendaten deutlich, dass eine Erweiterung des dreistufigen konzeptionellen Lebenszyklusmodells nach Byers und Braham (1948) notwendig ist: Als nulltes Stadium kann die Entwicklung konvektionsförderlicher, präkonvektiver Umgebungen angesehen werden. Neueste Satellitengenerationen können mit Hilfe von neuer Messtechnik hochaufgelöste horizontale und vertikale Temperatur- und Feuchteprofile der Troposphäre bestimmen[10], die beispielsweise Rückschlüsse auf bodennahe Feuchteflusskonvergenzen als Voraussetzung für konvektive Initiierung ermöglichen (z. B. Kalthoff et al., 2009). Das Cumulusstadium einer konvektiven Zelle kann nach den eingangs des Abschnitts genannten Studien mit Satellitendaten ferner in zwei Stufen unterteilt werden: Die erste Stufe, die frühe Wachstumsphase, entspricht der anfänglichen Intensivierung eines Aufwindbereichs. Sie endet, wenn die Rate der Temperaturabnahme an der Wolkenobergrenze abnimmt. Ihre beobachtete

[10] Siehe z. B.: `https://www.eumetsat.int/meteosat-third-generation`

Dauer hängt eng mit dem Auslösemechanismus sowie den Detektionskriterien der entsprechenden Algorithmen zusammen (Senf et al., 2015). Die zweite Stufe stellt eine erweiterte Wachstumsphase dar, deren Dauer im Allgemeinen zwischen 30 und 45 min liegt. Im erweiterten konzeptionellen Lebenszyklusmodell bleiben das Reife- sowie das Dissipationsstadium erhalten. Neben der Präzisierung der Beschreibung des Lebenszyklus einzelner konvektiver Zellen ermöglichen Satellitendaten auch eine gute Analyse der Entwicklung von MCS. Beispielsweise gehen horizontal weit ausgedehnte MCS mit einer großen Fläche von tiefen Temperaturen an der Wolkenobergrenze sowie einer erhöhten Anzahl von Blitzen zwischen Erdboden und Wolke einher (Mattos und Machado, 2011). Zudem korreliert die Lebensdauer dieser konvektiven Systeme mit ihrer horizontalen Ausdehnung (z. B. Feng et al., 2012).
Auch aus Daten von Niederschlagsradaren und den entsprechenden Zellverfolgungsalgorithmen (s. Kapitel 4.1) konnten Charakteristika der Lebenszyklen konvektiver Zellen identifiziert werden. MacKeen et al. (1999) verdeutlichten jedoch bereits, dass verschiedene radarbasierte Beobachtungsgrößen sowie Kombinationen von diesen eine niedrige Korrelation mit der Lebensdauer konvektiver Zellen aufweisen. Das Grundproblem ist ein großes Ungleichgewicht in der Anzahl von Zellen mit kurzer und langer Lebensdauer – von letzteren treten weitaus weniger auf. Die abgeleiteten Eigenschaften von Zellen mit kurzer und langer Lebensdauer unterscheiden sich nicht stark genug, um basierend auf der Statistik eine ausreichend scharfe Vorhersage zu treffen. Zudem dominieren kurzlebige konvektive Zellen die Evaluation. Ähnliche Häufigkeitsverteilungen der Lebensdauer zeigen sich in vielen weiteren Studien (z. B.. Wilson et al., 1998; Davini et al., 2012; Meyer et al., 2013; Wapler, 2021). Davini et al. (2012) zeigten für Zellen in Norditalien, dass diese ihre maximale Intensität (genauer: den größten Radarreflektivitätsfaktor; s. Kapitel 4.1.1) bereits in der ersten Hälfte ihres Lebenszyklus erreichen. Dies bestätigten Brisson et al. (2018) auch für eine größere

Stichprobe von simulierten Zellen in einem numerischen Atmosphärenmodell. Die größte flächenhafte Ausdehnung entsteht nach Davini et al. (2012) hingegen erst in der zweiten Hälfte des Lebenszyklus. Darüber hinaus weisen die Autoren darauf hin, dass die anfängliche Wachstumsrate der Zellfläche ein Indikator für die zu erwartende Lebensdauer sein könnte. Untersuchungen von Weusthoff und Hauf (2008) und Wapler (2021) zeigen, dass der Verlauf der flächenhaften Ausdehnung einer konvektiven Zelle im statistischen Mittel gut durch eine nach unten geöffnete Parabel oder eine halbe Sinusperiode approximiert werden kann. Gleichzeitig ist die Variabilität einzelner Lebenszyklen sehr hoch, welche auf eine hohe Vorhersageunsicherheit hindeutet (s. Kapitel 5.1 für detaillierte Informationen und Analysen hierzu).
Blitzdaten sind besonders als Indikator für die Intensivierung konvektiver Zellen nützlich: Steigt die Anzahl von Blitzen in Verbindung mit einer konvektiven Zelle plötzlich schnell an (*Lightning Jump*), ist mit einer Intensivierung einer Zelle innerhalb der nächsten 15 – 30 min zu rechnen (z. B. Mikuš Jurković et al., 2015; Wapler, 2017). Eine Konsequenz ist daher, dass konvektive Systeme, die mindestens einen sprunghaften Anstieg der Blitzanzahl aufweisen, eine längere Lebensdauer als solche ohne einen derartigen Anstieg haben (z. B. Chronis et al., 2015). Wie bereits während der theoretischen Beschreibungen der Lebenszyklen konvektiver Zellen (Kapitel 2.2.1 bis 2.2.4) und in Kapitel 2.3 deutlich wurde, spielen auch die atmosphärischen Bedingungen in der Umgebung konvektiver Zellen eine wichtige Rolle für deren Entstehung und deren Lebenszyklus durch verschiedene Prozesse und Wechselwirkungen.

In den letzten Jahren wurden daher zunehmend *Nowcasting*-Verfahren entwickelt und erweitert, die Daten aus verschiedenen Messmethoden und numerischen Vorhersagemodellen im Sinne des Multi-Daten-Ansatzes kombinieren, den schon MacKeen et al. (1999) vorschlugen (vgl. Kapitel 1; Schmid et al., 2019). Während solche Verfahren (insbesondere ihre internen Lebenszyklusmodelle) aufgrund der limitierten Verfügbarkeit

von Beobachtungsdaten konzeptionell zunächst recht einfach waren (z. B. Dixon und Wiener, 1993; Hand und Conway, 1995), hat ihre Komplexität in den letzten Jahren stark zugenommen. Mecikalski et al. (2015) zeigten, dass Verfahren aus der Statistik und dem maschinellen Lernen ein satellitenbasiertes Modell zur Erkennung der Auslösung von Konvektion durch die Berücksichtigung der latenten Instabilität in Form der CIN und der CAPE aus NWV-Vorhersagen signifikant verbessern können. Das System *Context and Scale Oriented Thunderstorm Satellite Predictors Development* (COALITION[11]) des Schweizer Wetterdienstes MeteoSchweiz schätzt den Verlauf des Lebenszyklus für die nächste Stunde probabilistisch auf der Basis bestimmter Blitz-, Radar-, Satelliten- und NWV-Daten ab und berücksichtigt zusätzlich den Einfluss der Orografie (Nisi et al., 2014). Schon bis zu 20 min im Voraus kann damit die Intensität einer konvektiven Zelle gut vorhergesagt werden. Die relative Wichtigkeit der Radar- und NWV-Daten ist dabei höher als die von anderen Datenquellen (Hamann et al., 2019). Die Multi-Daten-Analyse von Zöbisch et al. (2020) zeigte für ausgewählte Umgebungsvariablen aus der NWV einen statistischen Zusammenhang zwischen der Lebensdauer konvektiver Zellen und der Luftfeuchte sowie der latenten Instabilität (CAPE), überraschenderweise nicht jedoch mit der vertikalen Windscherung. Der Nutzen der mittels Fernerkundungsmethoden bestimmten Variablen ist ihren Untersuchungen zufolge größer als der Nutzen der von ihnen betrachteten atmosphärischen Umgebungsvariablen. Sie schlagen Untersuchungen mit weiteren Umgebungsvariablen und Kenngrößen vor, um potentiell besser geeignete Prädiktoren für *Nowcasting*-Verfahren zu identifizieren.

[11] `https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/warning-and-forecasting-systems/nowcasting.subpage.html/en/data/projects/2009/coalition.html`

Die Untersuchungen der vorliegenden Arbeit bewegen sich in diesem Bereich des Multi-Daten-Ansatzes. Unabhängig von der Organisationsform konvektiver Zellen wird untersucht, welche Zellattribute und welche Umgebungsvariablen für die Abschätzung des Lebenszyklus relevant sind. Dazu wird eine große Anzahl von sehr unterschiedlichen Umgebungsvariablen betrachtet. Statistische Verfahren, die sich für andere Fragestellungen des *Nowcastings* bewährt haben (z. B. Mecikalski et al., 2015; Czernecki et al., 2019), werden auf ihr Potential untersucht, das *Nowcasting* des Lebenszyklus konvektiver Zellen durch das Einbeziehen von Umgebungsvariablen aus NWV-Vorhersagen in einen Zellverfolgungsalgorithmus zu verbessern (s. Kapitel 3, 5 und 6). Damit liefert die vorliegende Arbeit einen wichtigen Beitrag im Rahmen der Optimierung von automatisierten Warnprozessen der Wetterdienste (vgl. Kapitel 1 und 7).

3 Methoden der Statistik und des maschinellen Lernens

Zur Datenaufbereitung, zur Analyse der Lebenszyklen konvektiver Zellen und der mit ihnen assoziierten Umgebungsbedingungen sowie zur Entwicklung von statistischen Vorhersageverfahren für verschiedene Eigenschaften konvektiver Zellen (Zellattribute) können prinzipiell viele unterschiedliche Methoden der Statistik und des maschinellen Lernens (*Machine Learnings*) verwendet werden. Die wichtigsten in der vorliegenden Arbeit genutzten Methoden werden in den folgenden Unterkapiteln vorgestellt. Die Untersuchung der Zusammenhänge zwischen unterschiedlichen Umgebungsvariablen basiert auf einer Korrelationsanalyse und einer Clusteranalyse (Kapitel 3.1 und 3.2).

Nach der Beschreibung dieser Methoden folgt eine Vorstellung von vier methodischen Ansätzen, die auf der Basis eines Datensatzes von Beobachtungsdaten eine statistische Vorhersage für verschiedene Attribute konvektiver Zellen liefern können. Dazu gehört die lineare Regression, welche den linearen Zusammenhang zwischen einer kontinuierlichen abhängigen Variablen und einer oder mehreren unabhängigen Variablen beschreibt (Kapitel 3.3.1). Es folgt eine Darstellung der logistischen Regression, welche als nicht-lineare Methode den Zusammenhang zwischen einer abhängigen und meist binären Variablen und einer oder mehreren unabhängigen Variablen untersucht (Kapitel 3.3.2). Eine Erweiterung der linearen Regression stellt ein polynomieller Ansatz dar, der die nicht-lineare Abhängigkeit in Form eines Polynoms höherer Ordnung beschreibt (Kapitel 3.3.3). Der sogenannte *Random Forest*, eine weitere Methode aus dem Bereich des maschinellen Lernens, bietet sowohl

die Möglichkeit der Vorhersage einer binären als auch einer kontinuierlichen abhängigen Variablen (Kapitel 3.4). Im Jargon des maschinellen Lernens modellieren Klassifikationsverfahren diskrete Variablen, während Regressionsverfahren kontinuierliche Variablen modellieren. Die logistische Regression zählt in dieser Definition – anders als ihr Name andeutet – folglich zu den Klassifikationsverfahren. In den jeweiligen Kapiteln wird die Quantifizierung des Einflusses der einzelnen unabhängigen Variablen ebenfalls thematisiert.

Kapitel 3.5 stellt Methoden vor, die der Aufbereitung von Datensätzen im Vorfeld der Anwendung eines statistischen Vorhersageverfahrens dienen. Kapitel 3 schließt mit der Einführung verschiedener Gütemaße, die eine Anwendung in der Analyse des Datensatzes und der Evaluation der statistischen Vorhersageverfahren finden (Kapitel 3.6).

3.1 Korrelations- und Hauptkomponentenanalyse

Die Korrelationsanalyse sowie die Hauptkomponentenanalyse können den Zusammenhang zwischen zwei Variablen $x^{(i)}$ und $x^{(k)}$ quantifizieren. Die Hauptkomponentenanalyse ist darüber hinaus auf hochdimensionale Problemstellungen anwendbar und dient allgemein als klassisches multivariates Verfahren auch zur Strukturierung und Reduzierung von Datensätzen (z. B. Wilks, 2006) oder zur Komplexitätsreduktion numerischer Modelle (z. B. Selten, 1995; Achatz und Schmitz, 1997; Wilhelm, 2014). In der vorliegenden Arbeit dient sie lediglich der Veranschaulichung von bivariaten Korrelationen, während die hauptsächliche Reduzierung des Datensatzes auf der Basis anderer Methoden und Zusammenhänge erfolgt (s. Kapitel 5.2.2 und 5.3.1).

3.1.1 Korrelationsanalyse

Der empirische Produkt-Moment-Korrelationskoeffizient nach Pearson r_P stellt ein Maß für den Grad des linearen Zusammenhangs zwischen zwei intervallskalierten Variablen $x^{(i)}$ und $x^{(k)}$ dar (z. B. Wilks, 2006):

$$r_P = \frac{cov_{x^{(i)},x^{(k)}}}{\sigma_{x^{(i)}}\sigma_{x^{(k)}}} \,. \tag{3.1}$$

Darin beschreibt $\sigma_{x^{(i)}}$ die empirische Standardabweichung der Variablen $x^{(i)}$ (analog $x^{(k)}$) gemäß

$$\sigma_{x^{(i)}} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(x_j^{(i)} - \overline{x}^{(i)}\right)^2} \tag{3.2}$$

und $cov_{x^{(i)},x^{(k)}}$ die empirische Kovarianz

$$cov_{x^{(i)},x^{(k)}} = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j^{(i)} - \overline{x}^{(i)}\right)\left(x_j^{(k)} - \overline{x}^{(k)}\right) \tag{3.3}$$

mit den empirischen Mittelwerten der Variablen $\overline{x}^{(i)}$ und $\overline{x}^{(k)}$ und der Stichprobengröße N. Weichen die Verteilungen der Variablen $x^{(i)}$ und $x^{(k)}$ zu stark von der Normalverteilung ab und/oder existiert ein nicht-linearer Zusammenhang zwischen den beiden, so liefert r_P trotz eines möglichen kausalen oder statistischen Zusammenhangs niedrige Werte.
In diesem Fall ist der empirische Rang-Korrelationskoeffizient nach Spearman besser geeignet (Spearman, 1904). Der Rang des j-ten Elements der Stichprobe $R\left(x_j^{(i)}\right)$ für die Variable $x^{(i)}$ entspricht der Position des Elements in der rangskalierten Reihe, d. h. das Element mit dem niedrigsten Variablenwert $x_{j_{min}}^{(i)}$ erhält Rang 1, also $R\left(x_{j_{min}}^{(i)}\right) = 1$, das mit dem zweitniedrigsten

Rang 2 etc. Damit folgt für den Rang-Korrelationskoeffizienten r_S:

$$r_S = \frac{cov_{R_{x^{(i)}},R_{x^{(k)}}}}{\sigma_{R_{x^{(i)}}}\sigma_{R_{x^{(k)}}}} \; . \tag{3.4}$$

Darin beschreibt $\sigma_{R_{x^{(i)}}}$ die empirische Standardabweichung des Rangs R der Variablen $x^{(i)}$ (analog $x^{(k)}$) gemäß

$$\sigma_{R_{x^{(i)}}} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left[R\left(x_j^{(i)}\right)-\overline{R}_{x^{(i)}}\right]^2} \tag{3.5}$$

und $cov_{R_{x^{(i)}},R_{x^{(k)}}}$ die empirische Kovarianz der Ränge

$$cov_{R_{x^{(i)}},R_{x^{(k)}}} = \frac{1}{N-1}\sum_{j=1}^{N}\left[R\left(x_j^{(i)}\right)-\overline{R}_{x^{(i)}}\right]\left[R\left(x_j^{(k)}\right)-\overline{R}_{x^{(k)}}\right] \tag{3.6}$$

mit den empirischen Mittelwerten der Ränge der Variablen $\overline{R}_{x^{(i)}}$ und $\overline{R}_{x^{(k)}}$. Diese sind jedoch einfach 0,5 $(N+1)$, was nach einigen Umformulierungen unter der Annahme, dass jeder Rang nur einmal angenommen wird, und wegen $\sigma_{R_{x^{(i)}}} = \sigma_{R_{x^{(k)}}}$ unter Verwendung der Summenformel für die Summe der ersten N natürlichen Quadratzahlen zu

$$r_S = 1 - \frac{6\sum_{j=1}^{N}\left[R\left(x_j^{(i)}\right)-R\left(x_j^{(k)}\right)\right]^2}{N\,(N-1)\,(N+1)} \tag{3.7}$$

führt. Besitzen mehrere Elemente einer Variablen denselben Rang, kann man Gleichung (3.7) benutzen, indem man den Mittelwert der jeweiligen Ränge als Rang der betroffenen Elemente verwendet. Allerdings ist die Übereinstimmung mit Gleichung (3.4) nicht exakt, da dann im Allgemeinen $\sigma_{R_{x^{(i)}}} \neq \sigma_{R_{x^{(k)}}}$ ist und die Summe aller quadrierten Ränge nicht mehr mit der Summe der natürlichen Quadratzahlen übereinstimmt.

Der Rang-Korrelationskoeffizient r_S ist folglich ein parameterfreies Maß für die Korrelation, da weder Linearität des Zusammenhangs noch eine Normalverteilung der Variablen vorliegen muss. Vielmehr kann er Korrelationen für beliebige monotone funktionale Zusammenhänge ohne Annahmen über die zugrunde liegenden Verteilungen bemessen. Ein weiterer Vorteil des Rang-Korrelationskoeffizienten r_S gegenüber r_P ist, dass er per Konstruktion robuster gegenüber Extremwerten und Ausreißern ist.
Zur Prüfung der Korrelationskoeffizienten und Untersuchung der statistischen Signifikanz des Zusammenhangs zwischen zwei Variablen ist ein zweiseitiger Einstichproben-t-Test dienlich (Student, 1908; Wilks, 2006). Ist der Wert der Prüfgröße

$$\phi = r\sqrt{\frac{N-2}{1-r^2}} \tag{3.8}$$

mit $r = r_P$ bzw. $r = r_S$ größer als der Wert der t-Verteilung mit einem Freiheitsgrad $f = N-2$ für ein bestimmtes Signifikanzniveau p (z. B. $p = 0{,}01$ oder 0,05), so ist die jeweilige Korrelation statistisch signifikant bezüglich dieses Niveaus. Je kleiner p gewählt wird, desto strenger ist die Testung. Eine Alternative ist die Betrachtung von Konfidenzintervallen, welche mit Hilfe der Fisher-Transformation bestimmt werden können und den Wertebereich von r abschätzen, der mit einer Wahrscheinlichkeit $w = 1-p$ den wahren Wert von r einschließt. w wird auch als Konfidenzniveau bezeichnet. Details hierzu finden sich beispielsweise in Kendall und Gibbons (1990) und Wilks (2006).

3.1.2 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse ist ein strukturentdeckendes statistisches Verfahren, in dem die Achsen eines n-dimensionalen Raums orthonormal transformiert werden, sodass diese in Richtung der zu den Eigenwerten der positiv semi-definiten, symmetrischen und diagonalisierbaren Kovarianzmatrix $\mathbf{C}$ ($n \times n$-Matrix) gehörigen Eigenvektoren zeigen (z. B. Preisendorfer, 1988).

Die Kovarianzmatrix setzt sich aus den einzelnen Beiträgen analog zu Gleichung (3.3) zusammen, d. h. $C_{i,k} = cov_{x^{(i)},x^{(k)}}$, was für $i = k$ mit $C_{i,k} = \sigma^2_{x^{(i)}}$ übereinstimmt.
Die Vorschrift zur Eigenwertzerlegung lautet

$$\mathbf{C} = \mathbf{E}\mathbf{L}\mathbf{E}^{-1} = \mathbf{E}\mathbf{L}\mathbf{E}^T \ . \tag{3.9}$$

Darin enthält die orthogonale $n \times n$-Matrix $\mathbf{E}$ spaltenweise die transformierten, orthonormalen Basisvektoren und $\mathbf{L} = \lambda \mathbb{1}$ mit der Einheitsmatrix $\mathbb{1}$ die entsprechenden Eigenwerte $\lambda = (\lambda_1, ..., \lambda_n)^T$. Die Matrix $\mathbf{L}$ stellt zugleich die Kovarianzmatrix im transformierten System dar. Im Allgemeinen ist es möglich, diese Hauptachsentransformation bezüglich einer beliebigen symmetrischen, positiv semi-definiten Metrik $\mathbf{M}$ durchzuführen, sodass gemäß Gleichung (3.9) folgende Eigenwertgleichung gilt:

$$\mathbf{CME} = \mathbf{EL} \ . \tag{3.10}$$

Aufgrund der positiven Semidefinitheit von $\mathbf{C}$ und $\mathbf{M}$ gilt für die Eigenwerte stets: $\lambda \geq 0$. Die Achse, entlang derer der Eigenvektor zum höchsten Eigenwert zeigt, wird als erste Hauptachse oder Hauptkomponente bezeichnet, welche zugleich den größten Anteil an der Gesamtstreuung abdeckt.
Zur Darstellung eines beliebigen Datenpunkts x_j durch die neue Basis betrachtet man die zentrierten ursprünglichen Koordinaten (Fluktuationen) $\mathbf{x}'_j = \mathbf{x}_j - \overline{\mathbf{x}} = (x'^{(1)}_j, ..., x'^{(n)}_j)^T$. Die $N \times n$-Matrix $\mathbf{X}'$ enthält zeilenweise die Fluktuationen von N Datenpunkten und die $N \times n$-Matrix A zeilenweise die Koordinaten der Datenpunkte $\mathbf{a}_j = (a^{(1)}_j, ..., a^{(n)}_j)^T$ im transformierten System. Damit lautet die Darstellung:

$$\mathbf{X}' = \mathbf{A}\mathbf{E}^T \ , \text{ bzw.}$$
$$\mathbf{x}'_j = \sum_{m=1}^{n} a^{(m)}_j \mathbf{e}_m \ . \tag{3.11}$$

Die Koordinaten eines beliebigen Datenpunkts $\mathbf{x}_j$ im transformierten System erhält man mittels orthogonaler Projektion der Fluktuationen auf die Eigenvektoren unter Berücksichtigung von $\mathbf{M}$. Aufgrund der Orthogonalität der Hauptkomponenten ($\mathbf{E}^T\mathbf{M}\mathbf{E} = \mathbb{1}$) lautet die Projektion wegen Gleichung (3.11):

$$\begin{aligned} \mathbf{A} &= \mathbf{X}'\mathbf{M}\mathbf{E} \ , \text{ bzw.} \\ a_j^{(m)} &= \mathbf{x}_j'^T \mathbf{M}\mathbf{e}_m \ . \end{aligned} \tag{3.12}$$

Darüber hinaus gilt wegen Gleichung (3.9) bzw. (3.10) im Standardfall $\mathbf{M} = \mathbb{1}$

$$\mathbf{C}\mathbf{x}_j' = \mathbf{E}\mathbf{L}\mathbf{E}^T\mathbf{x}_j' = \sum_{m=1}^{n} \lambda_m \left(\mathbf{x}_j' \cdot \mathbf{e}_m\right) \mathbf{e}_m \ . \tag{3.13}$$

Die paarweisen Korrelationen entlang der Hauptkomponenten sind per Konstruktion gleich Null (Preisendorfer, 1988). Die Varianz entlang des m-ten Eigenvektors $\mathbf{e}_m$ entspricht somit dem m-ten Eigenwert λ_m. Der Anteil dieser Varianz an der Gesamtstreuung σ_{tot} ist folglich:

$$\frac{\sigma_{\mathbf{e}_m}}{\sigma_{tot}} = \frac{\lambda_m}{\sum_{q=1}^{n} \lambda_q} \ . \tag{3.14}$$

Zur Eliminierung der Variablendimensionen und Normierung des Wertebereichs unterschiedlicher Variablen bietet es sich häufig an, vor der Hauptachsentransformation die Eingangsdaten über die Vorschrift

$$\tilde{x}_j^{(i)} = \frac{x_j^{(i)} - \overline{x}^{(i)}}{\sigma_{x^{(i)}}} \tag{3.15}$$

zu transformieren. Diese Transformation wird als Standardisierung oder z-Transformation bezeichnet und kombiniert eine Zentrierung und eine Skalierung der Daten. Es handelt sich dabei um eine monotone Transformation, welche die Ordnung der Elemente erhält. Das bedeutet, dass der Rang-Korrelationskoeffizient nach Spearman r_S zwischen zwei Variablen invariant bezüglich dieser Transformation ist. Darüber hinaus erhält sie die Schiefe der Verteilung der Werte von $x^{(i)}$.

Nach der z-Transformation ist $\tilde{\sigma}_{x^{(i)}} = 1$, die Kovarianzmatrix entspricht daher der Korrelationsmatrix und es gilt: $\tilde{\sigma}_{tot} = n$. Erscheint im Spezialfall $n = 2$ für $\mathbf{M} = \mathbb{1}$ im Streudiagramm die durch die beiden Hauptachsen aufgespannte 3σ-Ellipse näherungsweise als Kreis, so sind beide Achsen etwa gleichwertig und die lineare Korrelation der beiden Variablen ist gering. Im Fall einer idealen Normalverteilung liegen rund 98,9 % der Datenpunkte innerhalb der 3σ-Ellipse (Wang et al., 2015).

3.2 k-Medoids-Clustering

Clusterverfahren ordnen Datenpunkte eines n-dimensionalen Raums einer bestimmten Anzahl von Gruppen (N_C) zu und sind somit strukturentdeckende Verfahren. Die Gruppenanzahl N_C ist in der Regel a priori nicht bekannt. Die Zuordnung zu den Gruppen, die Cluster genannt werden, erfolgt aufgrund von Gemeinsamkeiten und Unterschieden der Datenpunkte (z. B. Wilks, 2006). Man unterscheidet hierarchische und nicht-hierarchische Clusterverfahren. Erstere erlauben während des Verfahrens keine neue Zuordnung von bereits einer Gruppe zugeordneten Datenpunkten, während letztere dies ermöglichen. Beide Verfahrenstypen benötigen jedoch eine Abstands- oder Dissimilationsmetrik $\mathscr{D}$, welche die Unterschiede zwischen Datenpunkten charakterisiert.

Ein häufig verwendetes nicht-hierarchisches Clusterverfahren ist das *k-Means*-Clustering (Lloyd, 1982[1]; MacQueen, 1967). Dieses teilt einen Datensatz dergestalt in N_C Cluster ein, dass die Summe der quadrierten Abweichungen der Datenpunkte von den Schwerpunkten der Cluster minimal ist, d. h. das Verfahren minimiert auch die Summe der Varianzen der Cluster. Dafür muss die Dimensionalität des Raums bekannt sein, um die Schwerpunkte zu definieren und die Abweichungen von ihnen zu berechnen. Ist allerdings einzig eine Dissimilationsmatrix $\mathbf{D}$ bekannt, welche die Abstände von N_O Objekten O zueinander beinhaltet, muss auf ein verwandtes Clusterverfahren, das sogenannte *k-Medoids*-Clustering zurückgegriffen werden (Kaufman und Rousseeuw, 1990). Der entwickelte Algorithmus wird auch als *Partitioning Around Medoids* (PAM) bezeichnet und konvergiert mit beliebigen Dissimilationsmetriken. Als Schwerpunkte der Cluster dienen hier vorhandene Objekte (Medoide). Das Verfahren minimiert die Summe der Abstände zwischen den Medoiden und den übrigen Clusterobjekten und ist damit robuster gegenüber Ausreißern als das *k-Means*-Clustering (z. B. Hastie et al., 2009).
Der sogenannte Silhouettenkoeffizient $\bar{s}$ beschreibt die Güte der Zuordnung durch das *k-Medoids*-Clustering (Rousseeuw, 1987). Dieser Koeffizient kann jeden Cluster einzeln ($\bar{s}_q$), oder alle Cluster gemeinsam ($\bar{S}$) als arithmetisches Mittel der einzelnen Silhouetten der Objekte bewerten:

$$\bar{s}_q = \frac{1}{N_O^{(q)}} \sum_{k=1}^{N_O^{(q)}} s_k^{(q)} \,, \tag{3.16}$$

$$\bar{S} = \frac{1}{N_C} \sum_{q=1}^{N_C} \bar{s}_q \,. \tag{3.17}$$

[1] Die Veröffentlichung in einer Zeitschrift erfolgte 1982, 25 Jahre nach der Verschriftlichung in einem Arbeitsbericht.

Darin ist $N_O^{(q)}$ die Anzahl der Objekte im q-ten Cluster. Die Silhouette des k-ten Objekts $O_k^{(q)}$ im Cluster C_q ist über

$$s_k^{(q)} = \begin{cases} 0 & \text{falls } N_O^{(q)} = 1 \ , \\ \frac{\overline{\mathscr{D}}\left(C_r, O_k^{(q)}\right) - \overline{\mathscr{D}}\left(C_q, O_k^{(q)}\right)}{\max\left[\overline{\mathscr{D}}\left(C_r, O_k^{(q)}\right), \ \overline{\mathscr{D}}\left(C_q, O_k^{(q)}\right)\right]} & \text{sonst} \end{cases} \tag{3.18}$$

definiert. Darin entspricht C_r demjenigen Cluster (der nicht C_q selbst ist), dessen Objekte dem Objekt $O_k^{(q)}$ im arithmetischen Mittel bezüglich der Dissimilationsmetrik $\mathscr{D}$ am nächsten liegen. Der Balken über $\overline{\mathscr{D}}$ kennzeichnet, dass in Gleichung (3.18) jeweils der arithmetische Mittelwert der Distanzen von $O_k^{(q)}$ zu allen im jeweiligen Cluster befindlichen Objekten (außer dem Objekt selbst) gemeint ist. Ist $s_k^{(q)} > 0$, so ist die Zuordnung zum Cluster C_q die beste Entscheidung, ist $s_k^{(q)} < 0$, so liegen die Objekte im Cluster C_r im Mittel näher an $O_k^{(q)}$ als diejenigen in Cluster C_q. Ist für einige Objekte $s_k^{(q)} < 0$, so ist N_C zu hoch oder zu niedrig gewählt. Ist für die meisten Objekte $s_k^{(q)} > 0$, so ist die Clusterkonfiguration eine gute Wahl. Eine starke (mittlere, schwache) Strukturierung liegt dann vor, wenn $s_k^{(q)} \in [0{,}75\,;\,1]$ ($[0{,}5\,;\,0{,}75)$, $[0{,}25\,;\,0{,}5)$) ist.

Eine grafische Darstellung des Clusterings zur einfacheren Interpretierbarkeit der Ergebnisse ist durch die Anwendung einer multidimensionalen Skalierung möglich (Pison et al., 1999). Dieses Verfahren schätzt die räumliche Konfiguration der Objekte O aus den paarweisen Distanzen ab, also die Positionierung der Objekte zueinander. Prinzipiell kann das Verfahren diese Konfiguration in hochdimensionalen Räumen bis zu einer maximalen Dimensionalität von $N_O - 1$ bestimmen. Meist findet jedoch eine Abbruchbedingung Verwendung, welche die Skalierung in einem möglichst niedrigdimensionalen Raum beendet, wenn dieser die Abstände der Objekte in sehr guter Näherung abbilden kann. Diese Bedingung begrenzt folglich die Dimensionalität des geschätzten Raums und erleichtert damit häufig die Interpretierbarkeit

der Ergebnisse. Die Darstellung der Objekte und der berechneten Cluster geschieht abschließend entlang der ersten beiden Hauptachsen des geschätzten Raums (vgl. Kapitel 3.1.2). Details zur multidimensionalen Skalierung finden sich z. B. in Backhaus et al. (2015).

3.3 Statistische Verfahren zur Vorhersage

Bei der Entwicklung von statistischen Vorhersageverfahren auf Basis eines (historischen) Datensatzes wird allgemein ein statistisches Modell erstellt, welches Schätzungen einer abhängigen Variablen anhand von bekannten oder ebenfalls durch ein Modell geschätzten unabhängigen Variablen liefert. Dazu ist es allgemein erforderlich, eine Aufspaltung des Datensatzes in zwei voneinander unabhängige Datensätze vorzunehmen. Einer der beiden Datensätze (Trainingsdatensatz) dient der Bestimmung der jeweiligen Modellparameter (Modellbildung), während anhand des anderen (Testdatensatz) die Vorhersagegüte des jeweiligen Modells evaluiert wird. Die in der vorliegenden Arbeit verwendeten Verfahren zählen zu den Methoden des überwachten Lernens (*Supervised Machine Learning*), bei denen in den Trainingsdatensätzen die abhängigen Variablen bekannt sind. Die folgenden Unterkapitel stellen den mathematischen Formalismus zur Modellbildung der verschiedenen statistischen Verfahren vor und erläutern Aspekte zur Evaluation und Interpretation.

3.3.1 Lineare Regression

Zusammenhänge zwischen einer kontinuierlichen abhängigen Variablen y und einer oder mehreren kontinuierlichen unabhängigen Variablen $\mathbf{x}$ werden nicht nur in der Meteorologie häufig auf der Basis einer (multiplen) linearen

Regression untersucht (z. B. Draper und Smith, 1998). Dieses strukturprüfende statistische Verfahren folgt dem Ansatz

$$y_j = \hat{y}(\mathbf{x}_j) + \varepsilon_j = b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)} + \varepsilon_j \; . \tag{3.19}$$

Der j-te von N Beobachtungswerten der Variablen y, die in diesem Ansatz auch Regressand oder Prädiktand heißt, wird durch einen Schätzwert $\hat{y}(\mathbf{x}_j)$ und ein Residuum ε_j dargestellt. Die Abschätzung von $\hat{y}(\mathbf{x}_j)$ erfolgt durch eine Linearkombination der N_x unabhängigen Variablen, welche hier als Regressoren bzw. Prädiktoren fungieren. Um die Modellparameter $\mathbf{b}$ dieses inversen Problems analytisch zu bestimmen, erfolgt eine globale Optimierung durch die Minimierung der Summe der quadratischen Fehler (Methode der kleinsten Quadrate). Dies entspricht in Matrixnotation der Minimierung der Kostenfunktion

$$J(\mathbf{b}) = ||\boldsymbol{\varepsilon}||^2 = ||\mathbf{y} - \hat{\mathbf{y}}||^2 = ||\mathbf{y} - \mathbf{X}\mathbf{b}||^2 \; , \tag{3.20}$$

in der $\mathbf{X}$ die $N \times (N_x + 1)$-Datenmatrix darstellt, welche in vielen Lehrbüchern den Namen Designmatrix trägt. Die doppelten Striche ($||\cdot||$) charakterisieren die Norm eines Vektors. In der Praxis ist es Usus, die euklidische Norm zu benutzen. Die erste Spalte von $\mathbf{X}$ enthält Einsen, während die übrigen Spalten die Werte der Prädiktoren $x_j^{(i)}$ für die N Beobachtungen beinhalten. $\mathbf{y}$ ist ein N-Spaltenvektor, der die bekannten Beobachtungswerte des Prädiktanden enthält, und b der $(N_x + 1)$-Spaltenvektor, der alle zu bestimmenden Modellparameter enthält. Die Minimierung führt auf die sogenannte Normalgleichung (z. B. Zeidler et al., 2012)

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{y} \tag{3.21}$$

mit der Lösung

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^{+} \mathbf{y} \; . \tag{3.22}$$

Darin bezeichnet $\mathbf{X}^+$ die (Moore-Penrose-)Pseudoinverse der Designmatrix. Sind die Spaltenvektoren von $\mathbf{X}^T\mathbf{X}$ linear unabhängig, d. h. $\text{rang}(\mathbf{X}^T\mathbf{X}) = N_x + 1$, so ist diese Matrix invertierbar und für Gleichung (3.22) existiert eine eindeutige Lösung. Dies ist demnach für linear unabhängige Prädiktoren der Fall. Numerisch wird **b** meist über das Cholesky-Verfahren ermittelt (Zeidler et al., 2012).

Die Modellparameter **b**, auch Regressionskoeffizienten genannt, haben eine besondere Bedeutung, da sie den marginalen Effekt der Änderung der Prädiktoren auf den Prädiktanden angeben (z. B. Backhaus et al., 2016). Wurden die Prädiktorwerte vor der Regression auf die gleiche Variationsskala gebracht, wie z. B. mit einer z-Transformation gemäß Gleichung (3.15), so kennzeichnen sie die relative Bedeutung der einzelnen Prädiktoren. Diese wird auch als Wichtigkeit der Prädiktoren bezeichnet (*Predictor Importance*). Im Fall unterschiedlicher Variationsskalen stellt der standardisierte Regressionskoeffizient $\tilde{b}^{(i)} = b^{(i)}\sigma_{x^{(i)}}\sigma_y^{-1}$ die Wichtigkeit des i-ten Prädiktors dar. Eine Normierung mit σ_y ist zum Vergleich der Wichtigkeit nicht zwingend notwendig. Zur Quantifizierung der Güte des Regressionsmodells sind verschiedene Maße nützlich. Eines davon ist der *Mean Squared Error* (*MSE*), welcher den mittleren quadratischen Fehler beschreibt und allgemein durch

$$MSE = \frac{1}{N'}\sum_{j=1}^{N}[y_j - \hat{y}(\mathbf{x}_j)]^2 = \frac{1}{N'}\sum_{j=1}^{N}\varepsilon_j^2 = \frac{1}{N'}SSE \tag{3.23}$$

mit $N' = N - N_x - 1$ und der Summe der quadratischen Fehler (*Sum of Squared Errors*; *SSE*) gegeben ist. Die Normierung mit $N - N_x - 1$ geht darauf zurück, dass man für die Fehlervarianz einen Bias-freien Schätzwert erhalten möchte (vgl. Wilks, 2006). Häufig findet auch der *Root Mean Squared Error* (*RMSE*) Anwendung, wobei gilt: $RMSE = \sqrt{MSE}$.

Gebräuchlich ist zudem das Bestimmtheitsmaß R^2, welches durch

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{3.24}$$

gegeben ist. Darin stehen die *Sum of Squares* (*total*; *SST*) und die *Regression Sum of Squares* (*SSR*) für die Summe der quadratischen Abweichungen der Beobachtungen bzw. der Schätzungen um das beobachtete arithmetische Mittel:

$$SST = \sum_{j=1}^{N} (y_j - \bar{y})^2 \tag{3.25}$$

$$SSR = \sum_{j=1}^{N} [\hat{y}(\mathbf{x}_j) - \bar{y}]^2 \ . \tag{3.26}$$

Das über den Parametersatz eindeutig definierte Modell kann im Anschluss auf einen unabhängigen Testdatensatz angewendet werden.

3.3.2 Logistische Regression

Das lineare Regressionsmodell aus Gleichung (3.19) ist generell auf binäre abhängige Variablen y anwendbar, d. h. mit einem Prädiktanden, der nur zwei verschiedene Werte annehmen kann (z. B. 0 und 1). Dies ruft jedoch einige Schwierigkeiten hervor. Die Schätzwerte des Prädiktanden $\hat{y}(\mathbf{x}_j)$ sind beispielsweise nicht begrenzt, und die Residuen ε_j sind im Allgemeinen nicht normalverteilt, da für sie gilt:

$$\varepsilon_j = 1 - B_j = 1 - \hat{y}(\mathbf{x}_j)^{y_j} \left[1 - \hat{y}(\mathbf{x}_j)\right]^{1-y_j} \ . \tag{3.27}$$

Die Abkürzung B_j steht darin für die Bernoulli-Verteilung $B\left[y_j \mid \hat{y}(\mathbf{x}_j)\right]$. Eine zentrale Annahme bei der Anwendung der Methode der kleinsten Quadrate ist jedoch eine Normalverteilung der Residuen (z. B. Wilks, 2006).

Die (multiple) logistische Regression bedient sich desselben linearen Ansatzes aus Gleichung (3.19) wie die lineare Regression, jedoch wird nicht die Variable $y \in \{0;1\}$ selbst als Prädiktand mit dem linearen Ansatz geschätzt. Stattdessen ermöglicht die sogenannte Logit-Transformation $\mathscr{L}$ eine Kopplung zwischen der (begrenzten) geschätzten Wahrscheinlichkeit $\hat{p}(y = 1 \mid \mathbf{x} = \mathbf{x}_j) \equiv \hat{p}(y_j = 1)$ und der (unbegrenzten) Linearkombination der unabhängigen Variablen $\mathbf{x}_j$ mit der Transformationsvorschrift (Hosmer und Lemeshow, 2000)

$$\mathscr{L}\hat{p}(y_j = 1) = \ln\left[\frac{\hat{p}(y_j = 1)}{1 - \hat{p}(y_j = 1)}\right] , \tag{3.28}$$

und dem Ansatz

$$\mathscr{L}\hat{p}(y_j = 1) = b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)} . \tag{3.29}$$

Dies entspricht gerade der Schätzung des natürlichen Logarithmus der Chance (*Odds*) O_j, welche dem Wahrscheinlichkeitsverhältnis der Eintrittswahrscheinlichkeit eines Ereignisses und deren Gegenwahrscheinlichkeit entspricht (vgl. Kapitel 3.6.1):

$$O_j = \frac{\hat{p}(y_j = 1)}{\hat{p}(y_j = 0)} = \frac{\hat{p}(y_j = 1)}{1 - \hat{p}(y_j = 1)} . \tag{3.30}$$

Aus den Gleichungen (3.19) und (3.28) lässt sich der neue Schätzwert $\hat{p}_j \equiv \hat{p}(y_j = 1)$ bestimmen:

$$\begin{aligned} \hat{p}_j &= \frac{\exp\left(b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right)}{1 + \exp\left(b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right)} \\ &= \frac{1}{2}\left[1 + \tanh\left(\frac{1}{2}\left\{b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right\}\right)\right] . \end{aligned} \tag{3.31}$$

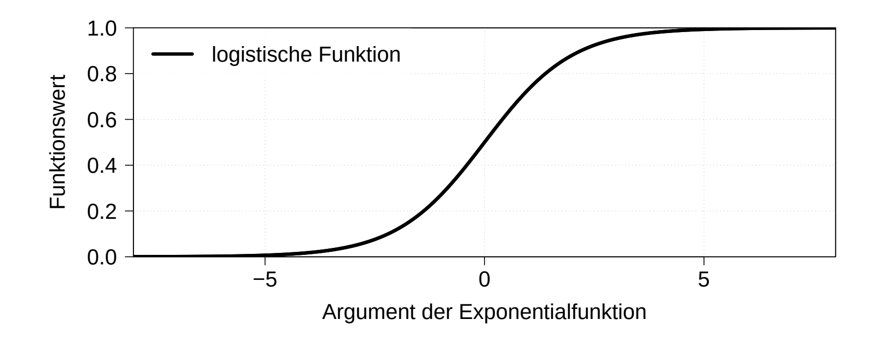


Abbildung 3.1: Exemplarischer Verlauf einer logistischen Funktion $f(s) = e^s\ (1+e^s)^{-1}$.

Eine solche Funktion gehört zur Klasse der Sigmoidfunktionen und wird auch als logistische Funktion bezeichnet (Abbildung 3.1; z. B. Backhaus et al., 2016). Der beobachtete Wert des Prädiktanden ist

$$p_j \equiv p(y_j = 1) = \delta_{y_j,1} = y_j \tag{3.32}$$

mit dem Kronecker-Delta

$$\delta_{p,q} = \begin{cases} 0 & \text{falls } p \neq q \\ 1 & \text{falls } p = q \end{cases} . \tag{3.33}$$

Die Residuen sind wiederum $\varepsilon_j = p_j - \hat{p}_j$, für die Gleichung (3.27) analog mit der Wahrscheinlichkeitsverteilung $B_j = B(p_j \mid \hat{p}_j)$ gilt. Unter Verwendung von Gleichung (3.31) folgt daraus:

$$\begin{aligned} \varepsilon_j &= 1 - B_j = 1 - \hat{p}_j^{p_j} \left[1 - \hat{p}_j\right]^{1-p_j} \\ &= \frac{\exp\left(b^{(0)} + \sum\limits_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right)^{1-p_j}}{1 + \exp\left(b^{(0)} + \sum\limits_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right)} . \end{aligned} \tag{3.34}$$

Auch diese Residuen sind im Allgemeinen nicht normalverteilt, sodass die Bestimmung der Modellparameter **b** mit der Methode der kleinsten Quadrate für die logistische Regression nicht geeignet ist. Stattdessen findet die sogenannte *Maximum Likelihood*-Methode Anwendung (z. B. Backhaus et al., 2016), die eine Maximierung der gemeinsamen Wahrscheinlichkeitsverteilung Λ (*Joint Likelihood*) bzw. deren Logarithmus L (*Log-Likelihood*) über alle N Datenpunkte anstrebt:

$$L = \ln(\Lambda) = \ln\left(\prod_{j=1}^{N} B_j\right) \quad . \tag{3.35}$$

Die Maximierung geschieht in der Praxis in der Regel iterativ.

Ist für eine Vorhersage a posteriori eine Transformation des Schätzwerts $\hat{p}_j = \hat{p}(y = 1 \mid \mathbf{x} = \mathbf{x}_j)$ in den ursprünglichen Wertebereich des Prädiktanden $\{0; 1\}$ erwünscht, so geschieht dies über die Vorschrift

$$\hat{y}(\mathbf{x}_j) = \begin{cases} 0 & \text{falls } \hat{p}_j < \mu \\ 1 & \text{falls } \hat{p}_j \geq \mu \end{cases} \quad . \tag{3.36}$$

Darin stellt $\mu \in [0; 1]$ einen Trennwert für die Wahrscheinlichkeit dar, den eine Vorhersage mindestens erreichen muss, damit die transformierte Vorhersage $\hat{y}(\mathbf{x}_j) = 1$ lautet. Im Folgenden wird μ daher als Entscheidungstrennwert (*Decision Threshold*) bezeichnet. In der praktischen Anwendung dient μ als Tuning-Parameter für binäre Vorhersagen basierend auf logistischen Regressionsmodellen (s. Kapitel 3.6.1 und 6.2.2).

Im Gegensatz zur linearen Regression ist ein Schluss von den Regressionskoeffizienten **b** bzw. $\tilde{\mathbf{b}}$ auf den marginalen Effekt oder die relative Wichtigkeit der Prädiktoren in der nicht-linearen logistischen Regression nicht direkt möglich. Stattdessen bedient man sich der Chance O_j aus Gleichung (3.30),

welche sich mit Gleichung (3.31) zu

$$O_j = \exp\left(b^{(0)} + \sum_{i=1}^{N_x} b^{(i)} x_j^{(i)}\right) \tag{3.37}$$

vereinfachen lässt. Eine Erhöhung des Werts des i-ten Prädiktors um eine Einheit führt darin offensichtlich zu einem zusätzlichen Faktor $\exp\left(b^{(i)}\right)$. Dieser wird als Effekt-Koeffizient E_K bezeichnet. Ist die Variationsskala der Prädiktoren unterschiedlich, so betrachtet man entsprechend den standardisierten Effekt-Koeffizienten $\tilde{E}_K = \exp\left(b^{(i)}\sigma_{x^{(i)}}\right) = E_K^{\sigma_{x^{(i)}}}$. Negative Werte von $b^{(i)}$ führen zu $0 < E_K < 1$. Die Wichtigkeit des Prädiktors kann in diesem Fall über den Kehrwert von E_K bzw. $\tilde{E}_K$ abgeschätzt werden.
Im Gegensatz zur linearen Regression und der Methode der kleinsten Quadrate ist die Berechnung eines Bestimmtheitsmaßes zur Quantifizierung der Güte des logistischen Regressionsmodells nicht möglich. In der Literatur finden sich viele Vorschläge für sogenannte Pseudo-Bestimmtheitsmaße, auch Pseudo-R^2 genannt, die auf Λ bzw. L, der Korrelation oder der erklärten Variation basieren (z. B. Veall und Zimmermann, 1996). Die Werte der verschiedenen Maße können innerhalb eines Modells stark variieren, sodass zur robusteren Einordnung der Güte eine kombinierte Betrachtung von mehreren Pseudo-Bestimmtheitsmaßen sinnvoll ist. Veall und Zimmermann (1996) zeigen, dass das Pseudo-R^2 von McKelvey und Zavoina (1975) als beste Approximation angesehen werden kann. Hosmer und Lemeshow (2000) hingegen betonen, dass sie die Verwendung der Pseudo-Bestimmtheitsmaße nicht empfehlen, da sie nicht wie das reguläre Bestimmtheitsmaß R^2 die Anpassungsgüte des Modells beurteilen. Weil die logistische Regression in der vorliegenden Arbeit nicht nur diagnostisch, sondern auch prognostisch Anwendung findet, werden keine Pseudo-Bestimmtheitsmaße berechnet. Stattdessen

kommen andere Gütemaße zur Evaluierung von binären Klassifikationsverfahren zum Einsatz, die auf dem Vergleich zwischen Vorhersage und Beobachtung basieren (s. Kapitel 3.6.1). Damit ist auch ein direkter Vergleich mit dem Vorhersageverfahren des *Random Forests* möglich (vgl. Kapitel 3.4).

3.3.3 Nicht-linearer Polynomansatz

Eine nicht-lineare Erweiterung des linearen Ansatzes in Gleichung (3.19) mit einem kontinuierlichen Prädiktanden y ist durch einen Polynomansatz der Ordnung N_p für N_x unabhängige Variablen $\mathbf{x}$ gegeben:

$$y_j = \hat{y}(\mathbf{x}_j) + \varepsilon_j = \sum_{i \in \mathbb{N}_0^{N_x}}^{(i_k)} b^{(i)} \mathbf{x}_j^{(i)} + \varepsilon_j \; . \tag{3.38}$$

Die Summe ist so zu verstehen, dass i alle N_x-Tupel (i_k) annehmen kann, für die gleichzeitig gilt:

$$\begin{aligned} \min[(i_k)] &\geq 0 \; , \\ \max[(i_k)] &\leq N_p \; , \\ \sum_{k=1}^{N_x} i_k &\leq N_p \; . \end{aligned}$$

Darin sei i_k der k-te Eintrag des i-ten N_x-Tupels (i_k). Außerdem sei der i-te Prädiktor durch

$$\mathbf{x}_j^{(i)} \equiv \prod_{m=1}^{N_x} \left(x_j^{(m)} \right)^{i_m} \tag{3.39}$$

dargestellt. Die Gesamtanzahl von Prädiktoren N_{po} ist in diesem Ansatz über den Binomialkoeffizienten

$$N_{po} = \binom{N_p + N_x}{N_x} - 1 \tag{3.40}$$

berechenbar. Für $N_p = 1$ ist dieser Ansatz identisch mit dem linearen Ansatz aus Gleichung (3.19), in dem $N_{po} = N_x$ ist. Für einen quadratischen Ansatz ($N_p = 2$) mit zwei unabhängigen Variablen $x^{(1)}$ und $x^{(2)}$ ($N_x = 2$) gilt beispielsweise:

$$(i_k) \in \{(0,0);\ (1,0);\ (0,1);\ (2,0);\ (0,2);\ (1,1)\}$$
$$\rightarrow N_{po} = \binom{4}{2} - 1 = 5\ , \tag{3.41}$$
$$\hat{y}(\mathbf{x}_j) = b^{(0,0)} + b^{(1,0)} x_j^{(1)} + b^{(0,1)} x_j^{(2)} + b^{(2,0)} \left(x_j^{(1)}\right)^2 + b^{(0,2)} \left(x_j^{(2)}\right)^2 + b^{(1,1)} x_j^{(1)} x_j^{(2)}\ . \tag{3.42}$$

Der nicht-lineare Ansatz bildet sowohl höhere Potenzen der einzelnen unabhängigen Variablen als auch kombinierte Mischterme ab. Daher sind in diesem Ansatz nicht-lineare Abhängigkeiten zwischen den Prädiktoren vorzufinden. Um potentielle Instabilitäten der Lösung dieses inversen Problems zu dämpfen, empfiehlt es sich, eine Regularisierung im Minimierungsverfahren anzuwenden (z. B. Nakamura und Potthast, 2015). Die sogenannte Tikhonov-Phillips-Regularisierung (Tikhonov, 1963; Phillips, 1962), meist nur Tikhonov-Regularisierung genannt, erweitert die Normalgleichung (3.21) auf

$$\left(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{P}^T\mathbf{P}\right)\mathbf{b} = \mathbf{X}^T\mathbf{y}\ . \tag{3.43}$$

Die Designmatrix $\mathbf{X}$ hat hier $N_{po} + 1$ Spalten und $\mathbf{b}$ ebenso viele Einträge. Diese Modifikation bedeutet, dass man die Methode der kleinsten Quadrate mit einer zusätzlichen Straffunktion (*Penalty Function*) anwendet:

$$J(\mathbf{b}) = ||\mathbf{y} - \mathbf{X}\mathbf{b}||^2 + \alpha||\mathbf{P}\mathbf{b}||^2\ . \tag{3.44}$$

Der Parameter $\alpha \geq 0$ wird Regularisierungsparameter genannt. Für $\alpha = 0$ entspricht Gleichung (3.44) der nicht regularisierten Kostenfunktion in Gleichung (3.20) für die ungedämpfte Lösung. Dominiert der Strafterm, so

ist man vom ursprünglichen Problem weit entfernt. Hansen (2010) und Nakamura und Potthast (2015) schlagen beispielsweise verschiedene Methoden zur optimalen Bestimmung von α vor, unter anderem basierend auf den Residuen $\boldsymbol{\varepsilon}$. Oft genügt es in der Praxis jedoch, verschiedene Werte zu testen. Die approximative Lösung für den Parametervektor $\mathbf{b}_\alpha$ ergibt sich analog zu Gleichung (3.22) über

$$\mathbf{b}_\alpha = \left(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{3.45}$$

bzw. im Standardfall $\mathbf{P} = \mathbb{1}$ über

$$\mathbf{b}_\alpha = \left(\mathbf{X}^T\mathbf{X} + \alpha\mathbb{1}\right)^{-1}\mathbf{X}^T\mathbf{y}\ . \tag{3.46}$$

Die Inverse von $\mathbf{X}^T\mathbf{X} + \alpha\mathbb{1}$ existiert für jede beliebige Designmatrix und beliebige Werte von $\alpha > 0$ (Nakamura und Potthast, 2015). Zur Vermeidung der numerischen Matrixinversion bietet sich eine Singulärwertzerlegung von $\mathbf{X}^T$ an (z. B. Zeidler et al., 2012). Ähnlich der Eigenwertzerlegung in Gleichung (3.9) lautet die Vorschrift zur Singulärwertzerlegung, die eine Verallgemeinerung zur Diagonalisierung nicht-quadratischer Matrizen darstellt,

$$\mathbf{X}^T = \mathbf{E}_L\mathbf{L}\mathbf{E}_R^* = \mathbf{E}_L\mathbf{L}\mathbf{E}_R^T\ , \tag{3.47}$$

worin der Stern (*) die Adjungierte einer Matrix kennzeichnet, die im Fall reellwertiger Einträge der Transponierten entspricht. $\mathbf{E}_L$ ist eine unitäre (hier: orthogonale) quadratische Matrix mit N Zeilen und Spalten, welche die sogenannten linken Singulärvektoren $\mathbf{e}_m^{(l)}$ enthält, $\mathbf{E}_R$ eine unitäre (hier: orthogonale) quadratische Matrix mit $N_{po}+1$ Zeilen und Spalten, die die rechten Singulärvektoren $\mathbf{e}_m^{(r)}$ enthält. $\mathbf{L}$ ist eine Diagonalmatrix, deren Einträge den Singulärwerten (erste $\mathrm{rang}(\mathbf{X})$ Diagonaleinträge) bzw. 0 (alle darauffolgenden Diagonaleinträge) entsprechen. Die Darstellung von $\mathbf{b}_\alpha$ bezüglich der Singulärbasis mit dem singulären System aus Singulärwerten sowie linken und

rechten Singulärvektoren $\left(\lambda_m, \mathbf{e}_m^{(l)}, \mathbf{e}_m^{(r)}\right)$ ist analog zu den Gleichungen (3.11) und (3.12)

$$\begin{aligned}\sum_{m=1}^{\text{rang}(\mathbf{X})} \left(\mathbf{b}_\alpha \cdot \mathbf{e}_m^{(r)}\right) \mathbf{e}_m^{(r)} &= \sum_{m=1}^{\text{rang}(\mathbf{X})} \left[\left(\mathbf{X}^T\mathbf{X} + \alpha\mathbb{1}\right)^{-1} \mathbf{X}^T \mathbf{y} \cdot \mathbf{e}_m^{(r)}\right] \mathbf{e}_m^{(r)} \\ &= \sum_{m=1}^{\text{rang}(\mathbf{X})} \lambda_m \left[\left(\mathbf{X}^T\mathbf{X} + \alpha\mathbb{1}\right)^{-1} \mathbf{y} \cdot \mathbf{e}_m^{(l)}\right] \mathbf{e}_m^{(r)} \ , \end{aligned} \tag{3.48}$$

da zudem in Analogie zu Gleichung (3.13)

$$\mathbf{X}^T \mathbf{y} = \sum_{m=1}^{\text{rang}(\mathbf{X})} \lambda_m \left(\mathbf{y} \cdot \mathbf{e}_m^{(l)}\right) \mathbf{e}_m^{(r)} \tag{3.49}$$

gilt und die Singulärvektoren orthonormale Basen bilden, d. h. unter anderem $\mathbf{e}_m^{(r)} \cdot \mathbf{e}_{m'}^{(r)} = \delta_{m,m'}$ gilt. Die Eigenwerte der Inversen der quadratischen Matrix $\mathbf{X}^T\mathbf{X}$ mit $N_{po} + 1$ Zeilen und Spalten sind zudem gleich den inversen Eigenwerten der Matrix, welche darüber hinaus mit dem Quadrat der Singulärwerte von $\mathbf{X}^T$ übereinstimmen (Zeidler et al., 2012). Somit muss elementweise gelten, dass die Projektion des Parametervektors auf den m-ten rechten Singulärvektor

$$\mathbf{b}_\alpha \cdot \mathbf{e}_m^{(r)} = \frac{\lambda_m}{\lambda_m^2 + \alpha} \left(\mathbf{y} \cdot \mathbf{e}_m^{(l)}\right) \tag{3.50}$$

ist. Die approximative Lösung für den Parametervektor ergibt sich daher schließlich zu

$$\mathbf{b}_\alpha = \sum_{m=1}^{\text{rang}(\mathbf{X})} \frac{\lambda_m}{\lambda_m^2 + \alpha} \left(\mathbf{y} \cdot \mathbf{e}_m^{(l)}\right) \mathbf{e}_m^{(r)} \ . \tag{3.51}$$

Für diesen Polynomansatz treten per Konstruktion starke Abhängigkeiten zwischen den Prädiktoren auf (Kollinearität), beispielsweise zwischen $x^{(i)}$ und $\left(x^{(i)}\right)^2$. Zur Untersuchung der relativen Wichtigkeit der unabhängigen Variablen empfiehlt es sich daher, zunächst den linearen Ansatz ungedämpft

über Gleichung (3.19) oder gedämpft über Gleichung (3.38) mit $N_p = 1$ anzuwenden. Für $N_p > 1$ ist die Interpretation einzelner Regressionskoeffizienten bei der Anwesenheit von Kollinearität nicht mehr klar, auch wenn diese den gemeinsamen Effekt nicht beeinflusst (Harrell, 2015). Für den Zweck der reinen Vorhersage mittels eines nicht-linearen Polynomansatzes spielt die Kollinearität demzufolge eine geringe Rolle, solange das Verfahren zur Schätzung der Regressionskoeffizienten stabil bleibt. Analog zur linearen Regression können zur Evaluation eines nicht-linearen Polynomansatzes prinzipiell die üblichen Maße wie der MSE bzw. $RMSE$ oder das Bestimmtheitsmaß R^2 in Gleichung (3.24) sowie weitere Gütemaße (s. Kapitel 3.6.2) herangezogen werden.

3.4 Der Random Forest

Random Forests als Vorhersageverfahren des maschinellen Lernens basieren auf sogenannten Entscheidungsbäumen. Um ihr Konzept zu verstehen, folgt zunächst eine Erläuterung der Idee und des mathematischen Formalismus für baumbasierte Methoden. Man unterscheidet Klassifikations- und Regressionsbäume. Erstere modellieren diskrete abhängige Variablen und letztere kontinuierliche abhängige Variablen. Der Formalismus der sogenannten CART-Methode (*Classification and Regression Trees*) wird zunächst anhand der Regressionsbäume in Anlehnung an Hastie et al. (2009), James et al. (2013), Kuhn und Johnson (2013) und Hatz (2018) vorgestellt.

3.4.1 Regressionsbäume

Baumbasierte Methoden teilen den N_{po}-dimensionalen Zustandsraum $\mathscr{X}$ der unabhängigen Variablen $\mathbf{x}$, welche in der Regel gleichzeitig die Prädiktoren darstellen, in N_m viele N_{po}-Hyperrechtecke (Orthotope) auf, welche im Folgenden die Abkürzung $\mathscr{R}^{(m)}$ erhalten. Für jedes der Orthotope wird ein konstanter Schätzwert des Prädiktanden $\hat{y}(\mathbf{x})$ bestimmt. Als nützlich erweist sich

dazu die Indikatorfunktion

$$I^{(m)}(\mathbf{x}) = I\left(\mathbf{x} \in \mathscr{R}^{(m)}\right) = \begin{cases} 0 & \text{falls } \mathbf{x} \notin \mathscr{R}^{(m)} \ , \\ 1 & \text{falls } \mathbf{x} \in \mathscr{R}^{(m)} \ . \end{cases} \tag{3.52}$$

Damit lässt sich ein einfaches Baummodell über

$$y_j = \hat{y}(\mathbf{x}_j) + \varepsilon_j = \sum_{m=1}^{N_m} b^{(m)} I^{(m)}(\mathbf{x}_j) + \varepsilon_j \tag{3.53}$$

mit den Konstanten $b^{(m)}$ definieren. Die Konstruktion der Orthotope wird durch rekursives binäres Aufteilen (Splitting) durchgeführt, d. h. zunächst wird der Zustandsraum in zwei Unterräume aufgeteilt, welche im Anschluss jeweils in zwei weitere Unterräume aufgeteilt werden und so weiter (Abbildung 3.2). Splits in mehr als zwei Unterräume (Multi-Splits) sind nicht üblich, da so die Anzahl von Datenpunkten in den Orthotopen zügig abnimmt und nur wenige Splitebenen entstehen. Darüber hinaus lässt sich jeder Multi-Split durch eine Verkettung von binären Splits darstellen. Die grundlegenden Freiheitsgrade in der Konstruktion eines Baums sind folglich (i) die Variablenauswahl für jeden Split s, (ii) die Wahl des Trennwerts t_s für den Split s und (iii) die Topologie des Baums, d. h. über wie viele Splitebenen ein Baum wachsen darf und wann kein weiterer Split vorgenommen wird (Abbruchbedingung).
Wie bei der linearen Regression in Kapitel 3.3.1 wird die Methode der kleinsten Quadrate zur theoretischen Bestimmung der optimalen Modellparameter **b** angewendet, hier auf die Kostenfunktion

$$J(\mathbf{b}) = ||\mathbf{y} - \mathbf{I}\mathbf{b}||^2 \ , \tag{3.54}$$

mit der $N \times N_m$-Indikatormatrix $\mathbf{I}$, deren Eintrag in der j-ten Zeile und m-ten Spalte gerade $I^{(m)}(\mathbf{x}_j)$ entspricht. Sie besteht folglich nur aus Nullen und Einsen, wobei jede Zeile aus einer Eins und $N_m - 1$ Nullen besteht,

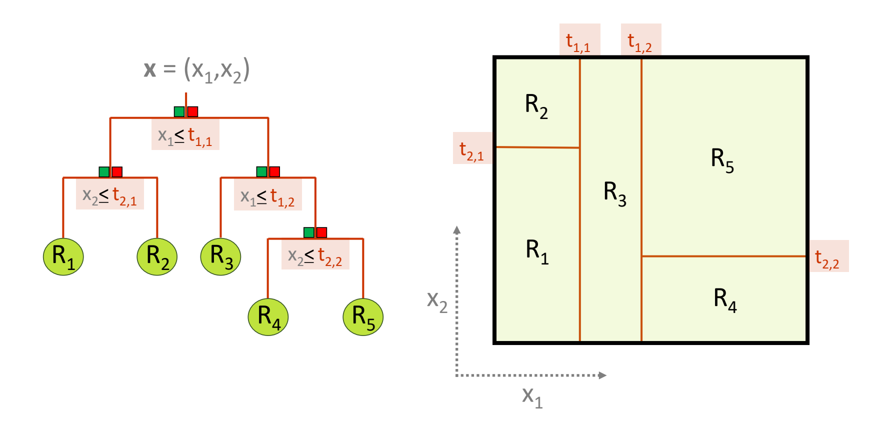


Abbildung 3.2: Illustration (a) eines beispielhaften Entscheidungsbaums mit $N_{po} = N_x = 2$ Prädiktoren und $N_m = 5$ Zwei-Orthotopen $\mathscr{R}^{(m)}$ über $N_s = 3$ Splitebenen (grün: Splitbedingung erfüllt; rot: nicht erfüllt) und (b) der entsprechenden Aufteilung im zweidimensionalen Zustandsraum mit den dazugehörigen Variablentrennwerten t_s. Nach Hastie et al. (2009).

und zählt daher zur Klasse der Indexmatrizen (z. B. Atanassov, 2014). Der Parametervektor **b** ergibt sich demnach zu

$$\mathbf{b} = \left(\mathbf{I}^T\mathbf{I}\right)^{-1}\mathbf{I}^T\mathbf{y} = \left[\operatorname{diag}\left(N_m\right)\right]^{-1}\mathbf{I}^T\mathbf{y} = \operatorname{diag}\left(N_m^{-1}\right)\mathbf{I}^T\mathbf{y}\,, \tag{3.55}$$

d. h. die Inverse von $\mathbf{I}^T\mathbf{I}$ ist stets wohldefiniert, weil $N_m > 0$ für alle m ist (s. u.). Da aber

$$\left(\mathbf{I}^\mathbf{T}\mathbf{y}\right)_m = \sum_{j=1}^{N} y_j I^{(m)}(\mathbf{x}_j) \tag{3.56}$$

gilt, folgt:

$$b^{(m)} = \frac{1}{N_m}\sum_{j=1}^{N} y_j I^{(m)}(\mathbf{x}_j) \equiv \overline{y}^{(m)}\,. \tag{3.57}$$

Für das m-te Orthotop ist der Modellparameter $b^{(m)}$ und somit auch der optimale Schätzwert des Prädiktanden $\hat{y}(\mathbf{x})$ mit $\mathbf{x} \in \mathscr{R}^{(m)}$ folglich einfach durch das arithmetische Mittel derjenigen Werte des Prädiktanden gegeben, die unter $\mathbf{x} \in \mathscr{R}^{(m)}$ beobachtet wurden.

Eine derartige Bestimmung ist in der Praxis jedoch nicht umsetzbar, da a priori die Orthotope $\mathscr{R}^{(m)}$ sowie deren Anzahl N_m nicht bekannt sind und ein Testverfahren über alle möglichen Realisierungen in der Regel geeignete Zeitskalen für die Rechenzeit weit überschreitet. Zwar existieren Vorschläge zur globalen Optimierung von Entscheidungsbäumen (z. B. Norouzi et al., 2015), jedoch basieren Algorithmen für *Random Forests* weitestgehend auf einer schrittweisen Minimierung für jeden Split. Dort wird jeweils nach der kleinsten Summe der quadratischen Abweichungen für alle $N_{po} = N_x$ Prädiktoren oder eine Auswahl von diesen gesucht (s. Kapitel 3.4.3). Die geringste Abweichung für den i-ten Prädiktor am s-ten Split für den Variablentrennwert t_s ist wegen Gleichung (3.57) durch

$$\begin{aligned} \varepsilon_s^{(i)}(t_s) = &\sum_{j=1}^{N} \left[y_j - \overline{y}_s^{(i,1)}(t_s) \right]^2 I_s^{(i,1)}(\mathbf{x}_j, t_s) \\ &+ \sum_{j=1}^{N} \left[y_j - \overline{y}_s^{(i,2)}(t_s) \right]^2 I_s^{(i,2)}(\mathbf{x}_j, t_s) \end{aligned} \tag{3.58}$$

gegeben, da bei jedem Split aus einem Unterraum $\mathscr{R}_s^{(0)}$ exakt zwei Unterräume $\mathscr{R}_s^{(i,1)}$ und $\mathscr{R}_s^{(i,2)}$ entstehen, die über

$$\mathscr{R}_s^{(i,1)}(t_s) = \left\{ \mathbf{x}_l \mid x_l^{(i)} \leq t_s \wedge \mathbf{x}_l \in \mathscr{R}_s^{(0)} \right\} , \tag{3.59}$$

$$\mathscr{R}_s^{(i,2)}(t_s) = \left\{ \mathbf{x}_l \mid x_l^{(i)} > t_s \wedge \mathbf{x}_l \in \mathscr{R}_s^{(0)} \right\} \tag{3.60}$$

definiert sind. Darin steht $\mathbf{x}_l$ für jeden möglichen Zustand, während $\mathbf{x}_j$ einen im Datensatz tatsächlich angenommenen Zustand bezeichnet. Durch das Austesten endlich vieler Werte $x^{(i)}$ für den Variablentrennwert t_s wird für

alle ausgewählten Prädiktoren der geringste Wert von $\varepsilon_s^{(i)}(t_s)$ bestimmt. Dafür sind maximal so viele Werte für t_s zu testen wie Datenpunkte im aufzuteilenden Unterraum $\mathscr{R}_s^{(0)}$ vorhanden sind. Anschließend erfolgt ein Vergleich der Werte der so verbleibenden (maximal N_x) Summen der Abweichungen und schließlich die Auswahl derjenigen Variablen mit dem entsprechenden optimalen Variablentrennwert, die für diesen Split die geringste Summe der Abweichungen aufweist. Damit sind die oben erwähnten Freiheitsgrade (i) und (ii) festgelegt.

Die Topologie des Baums (iii) ergibt sich durch die Forderung einer minimalen Anzahl von Datenpunkten innerhalb eines jeden Orthotops. Sind in einem Unterraum nach einer bestimmten Anzahl von Splits weniger als N_{min} Datenpunkte vorzufinden, so wird dieser nicht mehr weiter gesplittet. Dieses Orthotop bezeichnet man auch als ein Blatt des Entscheidungsbaums. Darüber hinaus ist es möglich einen Entscheidungsbaum zu stutzen, d. h. die Anzahl von Splitebenen a posteriori zu verringern, da es bei kleinen Werten für N_{min} zu einer Überanpassung (*Overfitting*) kommen kann. Beim sogenannten Kosten-Komplexität-Kriterium wird die Summe der quadratischen Abweichungen in den Blättern der Anzahl von Blättern eines gestutzten Baums N_m (bzw. βN_m mit $\beta \geq 0$ als Tuningparameter) gegenübergestellt, sodass die Abweichungen sowie die Anzahl von Blättern möglichst gering sind. Durch den Vergleich verschiedener Bäume, die unterschiedlich gestutzt werden, findet man damit adaptiv einen optimal gestutzten Entscheidungsbaum, der weniger überangepasst ist als der vollständig ausgewachsene Baum. Details hierzu finden sich beispielsweise in Breiman et al. (1984) oder Hastie et al. (2009). Für $N_{min} \not\ll N$ kommt es hingegen in der Regel zu einer Unteranpassung des klein gewachsenen Entscheidungsbaums, da relevante Strukturen verborgen bleiben.

3.4.2 Klassifikationsbäume

Klassifikationsbäume entsprechen konzeptionell den Regressionsbäumen, unterscheiden sich jedoch in der Formulierung der Kriterien für das Splitting und das Stutzen der Bäume. Wie in Kapitel 3.3.2 erläutert, ist die Methode der kleinsten Quadrate für binäre Klassifikationsprobleme (und ebenso für multikategorische Probleme mit N_K Klassen der abhängigen Variablen) nicht geeignet, sodass auch Gleichung (3.58) für das Splitting in Klassifikationsbäumen nicht angewendet wird. Mit $m = \{1,2\}$ nimmt in den Unterräumen $\mathscr{R}_s^{(i,m)}$ der Prädiktand genau $n_k^{(i,m)}$-mal den Wert der k-ten Klasse υ_k an, d. h. deren Anteil beläuft sich auf

$$f^{(i,m,k)} \equiv f^{(i,m)}(y = \upsilon_k) = \frac{n_k^{(i,m)}}{N^{(i,m)}} \ , \tag{3.61}$$

mit der Gesamtanzahl von Datenpunkten in jedem Unterraum $N^{(i,m)}$. Als Schätzwert des Prädiktanden $\hat{y}(\mathbf{x})$ mit $\mathbf{x} \in \mathscr{R}_s^{(i,m)}$ in beiden Unterräumen wird der jeweils am häufigsten auftretende Wert

$$\upsilon_{max}^{(i,m)} = \left\{ \upsilon_k \mid k = k_{max}^{(i,m)} = \text{argmax}\left(f^{(i,m,k)} \right) \right\} \tag{3.62}$$

verwendet. Zur Beschreibung der Abweichungen in den Unterräumen haben sich drei verschiedene Kenngrößen etabliert: der Missklassifikationsfehler (MF), der Gini Index (GI) und die Kreuz-Entropie (KE), welche häufig auch als Devianz bezeichnet wird. Sie berechnen sich in jedem Unterraum gemäß

$$MF^{(i,m)} = 1 - f^{(i,m,k_{max}^{(i,m)})} \ , \tag{3.63}$$

$$GI^{(i,m)} = \sum_{k=1}^{N_k} f^{(i,m,k)} \left(1 - f^{(i,m,k)} \right) \ , \tag{3.64}$$

$$KE^{(i,m)} = \sum_{k=1}^{N_k} f^{(i,m,k)} \ln \left(f^{(i,m,k)} \right) \ . \tag{3.65}$$

Hastie et al. (2009) empfehlen die Verwendung des GI oder der KE, da diese sensitiver in Bezug auf die Klassenanteile in der Unterräumen reagieren als der MF. Das Minimierungsproblem aus Gleichung (3.58) wird daher bei Klassifikationsbäumen beispielsweise durch die Minimierung der Summe der Gini Indizes in den beiden Unterräumen des Splits ersetzt:

$$\begin{aligned} \varepsilon_s^{(i)}(t_s) &= \sum_{j=1}^{N} GI^{(i,1)} I_s^{(i,1)}(\mathbf{x}_j, t_s) + \sum_{j=1}^{N} GI^{(i,2)} I_s^{(i,2)}(\mathbf{x}_j, t_s) \\ &= \sum_{m=1}^{2} N^{(i,m)} GI^{(i,m)} \ . \end{aligned} \tag{3.66}$$

Analog wird im Kosten-Komplexität-Kriterium zur Bestimmung der optimalen Topologie des Entscheidungsbaums eines der drei gelisteten Maße benutzt.

3.4.3 Der Random Forest als Kombination von Entscheidungsbäumen

Random Forests stellen eine Menge aus dekorrelierten Entscheidungsbäumen dar (Breiman, 2001). Die Vorhersagen von *Random Forests* ergeben sich mittels der Methode des sogenannten *Bootstrap Aggregatings* (kurz: *Baggings*; Breiman, 1996). Beim *Bagging* im *Random Forest* wird der vorliegende Datensatz durch Ziehen mit Zurücklegen von N_{bag} Datenpunkten in N_{Baum} Datensätze aufgeteilt. Aus jedem dieser Datensätze wird ein eigenständiger Entscheidungsbaum gebildet. Am Ende des Verfahrens erfolgt eine Kombination der resultierenden Schätzwerte (s. u.).

Das *Bagging* eignet sich allgemein besonders für Methoden mit hoher Varianz und niedrigem Bias (Hastie et al., 2009). Ungestutzte Entscheidungsbäume, welche per Konstruktion komplexe Strukturen in den Daten erfassen können, weisen bei niedrigem N_{min} aufgrund der Überanpassung an den verwendeten Datensatz einen sehr niedrigen Bias auf. Andererseits ist eine hohe Varianz

aufgrund der Überanpassung vorbestimmt. Darüber hinaus führen kleine Modifikationen im Datensatz zu anderen Entscheidungen für das Splitting über Gleichung (3.58) bzw. (3.66), was in sehr unterschiedlichen Bäumen resultieren kann. Solche Modelle bezeichnen viele Autoren auch als schwache Lerner. Durch das *Bagging* wird die Stabilität und Genauigkeit solcher schwachen Lerner verbessert sowie die Varianz und die Überanpassung reduziert. Die Varianz des kompletten *Random Forests* hängt dabei multiplikativ von der Varianz der einzelnen Entscheidungsbäume sowie deren paarweisen Korrelationen ab, sofern der *Random Forest* hinreichend groß ist. Details dazu finden sich in Hastie et al. (2009). Ziel des *Baggings* ist es daher, die Bäume zu dekorrelieren und gleichzeitig die Varianz der Bäume dadurch nicht zu stark zu erhöhen.

Die Dekorrelation der Entscheidungsbäume wird dadurch erreicht, dass bei jedem Split in den Bäumen lediglich eine bestimmte Anzahl $N_{split} \leq N_{po}$ von Prädiktoren betrachtet wird. Diese ergibt sich durch zufälliges Ziehen aus allen Prädiktoren ohne Zurücklegen. Die bestmögliche Wahl für den Parameter N_{split} ist in der Regel von der Gesamtzahl der Prädiktoren, deren Korrelation, der Problemstellung sowie den zugrundeliegenden Daten abhängig (z. B. Hastie et al., 2009; Bernard et al., 2009). Die typischen Standardwerte sind $N_{split} = \lfloor N_{po}/3 \rfloor$ für Regressions- sowie $N_{split} = \lfloor \sqrt{N_{po}} \rfloor$ für Klassifikationsprobleme. Dabei steht die Klammer ($\lfloor \cdot \rfloor$) für die Gaußklammer, die eine Abrundung auf die nächstkleinere ganze Zahl kennzeichnet. Bei sehr kleinen Werten für N_{po} muss N_{split} in der Regel größere Werte als die jeweiligen Standardwerte annehmen, um die bestmögliche Vorhersagegüte zu erreichen. Auch wenn nicht garantiert ist, dass diese Wahl für N_{split} in jedem Fall die bestmögliche ist, so stellt sie zumindest einen guten Richtwert dar. Als Maß für die Abweichungen beim Splitting wird für Regressionsbäume der *MSE* wie in Gleichung (3.58), für Klassifikationsbäume meist der Gini-Index *GI* wie in Gleichung (3.66) verwendet. Letzterer spart gegenüber der Kreuz-Entropie *KE* etwas an Rechenzeit.

Darüber hinaus wird bei der Verwendung des *Random Forests* für Regressionsprobleme für jeden Entscheidungsbaum die Abbruchbedingung $N_{min} = 5$ und für Klassifikationsprobleme $N_{min} = 1$ als Standardwert empfohlen. Auch wenn Segal (2004) zeigte, dass durch die Beschränkung der Anzahl von Splitebenen kleine Verbesserungen bezüglich der Rechenzeit bei der Erstellung eines *Random Forests* auftreten können, konstatieren Hastie et al. (2009), dass die Generierung eines vollständig ausgewachsenen Baums mit den genannten Abbruchbedingungen meist eines nur wenig höheren Rechenaufwands bedarf. Gleichzeitig eliminiert man damit einen Tuning-Parameter.

Die resultierenden Schätzwerte aus den einzelnen N_{Baum} Entscheidungsbäumen $\hat{y}^{(q)}(\mathbf{x})$ kombiniert ergeben einen gemeinsamen Schätzwert des *Random Forests*. Für Regressionsbäume bestimmt das arithmetische Mittel den kombinierten Schätzwert:

$$\begin{aligned}\hat{y}(\mathbf{x}_j) &= \frac{1}{N_{Baum}} \sum_{q=1}^{N_{Baum}} \hat{y}^{(q)}(\mathbf{x}_j) \overset{(3.53)}{=} \frac{1}{N_{Baum}} \sum_{q=1}^{N_{Baum}} \sum_{m=1}^{N^{(m,q)}} b^{(m,q)} I^{(m,q)}(\mathbf{x}_j) \\ &\overset{(3.57)}{=} \frac{1}{N_{Baum}} \sum_{q=1}^{N_{Baum}} \sum_{m=1}^{N^{(m,q)}} \overline{y}^{(m,q)} I^{(m,q)}(\mathbf{x}_j) \; . \end{aligned} \tag{3.67}$$

Für Klassifikationsbäume betrachtet man das sogenannte *Cutoff*-Verhältnis

$$\begin{aligned} C_v^{(k)}(\mathbf{x}_j) &= \frac{1}{C^{(k)}} \mathscr{T}_q^{(k)} \left[\sum_{m=1}^{N^{(m,q)}} b^{(m,q)} I^{(m,q)}(\mathbf{x}_j) \right] \\ &\overset{(3.62)}{=} \frac{1}{C^{(k)}} \mathscr{T}_q^{(k)} \left[\sum_{m=1}^{N^{(m,q)}} \overline{\upsilon}_{max}^{(m,q)} I^{(m,q)}(\mathbf{x}_j) \right] \\ &\equiv \frac{F^{(k)}(\mathbf{x}_j)}{C^{(k)}} \; . \end{aligned} \tag{3.68}$$

Darin steht $\overline{\upsilon}_{max}^{(m,q)}$ für den Schätzwert des Prädiktanden im m-ten Blatt des q-ten Entscheidungsbaums. Der Operator $\mathscr{T}_q^{(k)}$ zählt, wie häufig $b^{(m,q)} = \overline{\upsilon}_{max}^{(m,q)}$ dem Wert der k-ten Klasse von den N_{Baum} einzelnen Schätzwerten der Entscheidungsbäume entspricht und teilt diese Zahl durch N_{Baum}. Er gibt demnach die relative Häufigkeit der Vorhersagen der Werte der k-ten Klasse $F^{(k)}(\mathbf{x}_j)$ im *Random Forest* an. $C^{(k)}$ ist der der k-ten Klasse zugeordnete, a priori festgelegte *Cutoff*-Wert, und es muss gelten:

$$\sum_{k=1}^{N_k} C^{(k)} = 1 \, . \tag{3.69}$$

Der kombinierte Schätzwert bestimmt sich schließlich über

$$\hat{y}(\mathbf{x}_j) = \left\{ \upsilon_k \mid k = \text{argmax} \left(C_v^{(k)}(\mathbf{x}_j) \right) \right\} \, . \tag{3.70}$$

Er entspricht demzufolge dem Wert der Klasse, die $C_v^{(k)}(\mathbf{x}_j)$ maximiert. Wählt man beispielsweise $C^{(k)} = N_k^{-1} \; \forall \, k$, so entspricht der kombinierte Schätzwert dem Wert derjenigen Klasse, die am häufigsten von den einzelnen Entscheidungsbäumen vorhergesagt wird. Im Fall $N_k = 2$ entspricht dies der (einfachen) Mehrheitsentscheidung. Wählt man hingegen für $N_k = 2$ beispielsweise $\mathbf{C} = (0{,}7\,; 0{,}3)$, so müssen 70 % der Entscheidungsbäume im *Random Forest* für die erste Klasse stimmen, damit die Vorhersage des *Random Forests* ebenfalls Klasse Eins lautet. Die Klasse Zwei wird hingegen schon vorhergesagt, wenn nur 30 % der Bäume diese Vorhersage stützen. Für binäre Prädiktanden mit $\upsilon \in \{0\,; 1\}$ ist die Bestimmung von $\hat{y}(\mathbf{x}_j)$ über das *Cutoff*-Verhältnis gemäß Gleichung (3.70) mit derjenigen, die in Gleichung (3.36) für die logistische Regression verwendet wurde, unter Verwendung folgender Äquivalenzen identisch:

$$\mu \equiv C^{(2)} \, , \tag{3.71}$$

$$\hat{p}_j = \hat{p}(y = 1 \mid \mathbf{x} = \mathbf{x}_j) \equiv F^{(2)}(\mathbf{x}_j) \, . \tag{3.72}$$

Ein *Random Forest* kann nicht zwischen dem marginalen Effekt und der Wichtigkeit der Prädiktoren unterscheiden, da das Splitting der Entscheidungsbäume skaleninvariant unter monotonen Transformationen ist. Es existieren mehrere Maße, welche die relative Wichtigkeit der Prädiktoren quantifizieren. Das einfachste Maß zählt, wie häufig ein Prädiktor für einen Split ausgewählt wurde. Dieser wurde dort gerade deswegen ausgewählt, weil er die Abweichungen bei einem Split in Gleichung (3.58) bzw. (3.66) minimiert und damit potentiell ein hohes Unterscheidungsvermögen in Bezug auf den Prädiktanden besitzt (Hatz, 2018).

Gebräuchlicher ist es, die Verbesserungen des Splitkriteriums ausgedrückt durch den MF, den GI oder die KE durch einen Prädiktor für alle Splits in den Entscheidungsbäumen zu bestimmen (Breiman et al., 1984; Breiman, 2001). Genauer formuliert: Für jeden Split wird beispielsweise der GI aus dem Unterraum $\mathscr{R}_s^{(0)}$ mit der mit $N^{(i,1)}N^{(0)-1}$ bzw. $N^{(i,2)}N^{(0)-1}$ gewichteten Summe der beiden GI aus den Unterräumen $\mathscr{R}_s^{(i,1)}$ und $\mathscr{R}_s^{(i,2)}$ verglichen und die Differenz als GI-Verringerung bezeichnet. Je höher das Mittel über alle Bäume der N_{Baum} Summen der mit dem jeweiligen $N^{(0)}$ gewichteten GI-Verringerungen aller Entscheidungsbäume ist, desto höher ist die Wichtigkeit des Prädiktors (Gini-Wichtigkeit).

Eine dritte Möglichkeit zur Bestimmung der Wichtigkeit eines Prädiktors ist, dessen Werte in den jeweiligen Datensätzen der Entscheidungsbäume zufällig zu permutieren, d. h. jedem Datenpunkt wird zufällig der Wert des Prädiktors eines anderen Datenpunkts zugeordnet, während die Werte der übrigen Prädiktoren und die des Prädiktanden des Datenpunkts konstant bleiben (Breiman, 2000). Damit wird ein zweiter *Random Forest* generiert, bei dem (wie für den ursprünglichen *Random Forest*) für jeden einzelnen Baum ein bestimmtes Fehlermaß, der sogenannte OOB-Fehler (s. u.), bestimmt wird. Im Anschluss bestimmt man die Differenz der N_{Baum} OOB-Fehler und mittelt diese anschließend. Je höher die mittlere Differenz, desto höher die

Wichtigkeit des Prädiktors (Strobl et al., 2007; Hastie et al., 2009). Dieses Maß wird auch als Permutations-Wichtigkeit bezeichnet.

Durch das *Bagging*, bei dem eine zufällige Ziehung von N_{bag} Datenpunkten aus dem kompletten Datensatz mit Zurücklegen erfolgt, gehen nicht alle Datenpunkte in die Bildung des q-ten Entscheidungsbaums ein. Diese sind – wörtlich übersetzt – außerhalb des Sacks (*Out of Bag*; OOB) für diesen Baum. Daher ist für diese Datenpunkte direkt bei der Modellbildung eine Abschätzung der Residuen $\varepsilon_j = y_j - \hat{y}(\mathbf{x}_j)$ und weiterer Fehlermaße in jedem Baum möglich (z. B. Hastie et al., 2009). Stabilisieren sich diese, so ist eine Erweiterung des *Random Forests* um weitere Bäume nicht notwendig. Tatsächlich geht der OOB-Fehler, den verschiedene Fehlermaße beschreiben und quantifizieren können, für eine genügend große Anzahl von Bäumen in den Fehler über, den man durch einen zweiten, ähnlich großen unabhängigen Datensatz erhalten würde. Prinzipiell ist somit keine Unterscheidung in Trainings- und Testdatensatz und auch keine Kreuzvalidierung nötig (Breiman, 2001; James et al., 2013). Zum Vergleich mit einer anderen Vorhersagemethode wie z. B. der logistischen oder (nicht-)linearen Regression ist es allerdings empfehlenswert, wie bei der anderen Methode eine Untersuchung mit mehreren *Random Forest*-Modellen mit jeweils denselben Trainings- und Testdatensätzen durchzuführen, da diese dort zur unabhängigen Quantifizierung der Vorhersagegüte notwendig ist (vgl. Kapitel 3.3).

3.5 Methoden zur Aufbereitung der Datensätze

Bevor ein Datensatz zur Modellbildung eines statistischen Vorhersageverfahrens verwendet werden kann, sind häufig einige vorbereitende Bearbeitungsschritte notwendig (*Preprocessing*; s. Kapitel 6.1.1). Die folgenden Kapitel

stellen dazu einige Grundlagen für bestimmte mathematische Transformationen der Daten (Kapitel 3.5.1) sowie Methoden des sogenannten *Resamplings* vor (Kapitel 3.5.2), die in der Datenvorbehandlung für die Modellstudien in Kapitel 6 Verwendung finden.

3.5.1 Mathematische Transformationen

Wilks (2006) führt an, dass wichtige Merkmale einer Variablen im ursprünglichen Variationsbereich verborgen bleiben können. Eine mathematische Transformation der Variablenwerte eines Datensatzes vor der Anwendung eines statistischen Verfahrens kann nützlich sein, um dessen Aussagekraft zu steigern.
Eine monotone Transformation, welche eine Variable $x^{(i)}$ entdimensionalisiert, zentriert und auf einen Variationsbereich mit der Standardabweichung $\sigma_{x^{(i)}} = 1$ normiert, ist die in Gleichung (3.15) bereits beschriebene z-Transformation. Diese erhält unter anderem die Schiefe der Verteilung (vgl. Kapitel 3.1.2). Werden die Werte der Prädiktoren für statistische Verfahren im Vorfeld z-transformiert, vereinfacht dies die Quantifizierung der Wichtigkeit der Prädiktoren in vielen Verfahren wie beispielsweise über die Modellparameter **b** bei der linearen Regression (vgl. Kapitel 3.3.1) oder den Effekt-Koeffizienten E_K bei der logistischen Regression (vgl. Kapitel 3.3.2). Bis auf die Tatsache, dass die Werte der Prädiktoren nach der Transformation physikalisch schwieriger zu interpretieren sind, hat eine solche Transformation keine Nachteile und keinen Effekt auf die Vorhersagen eines statistischen Vorhersageverfahrens und deren Güte sowie auf die Wichtigkeit der Prädiktoren (z. B. Kuhn und Johnson, 2013).
Eine für unimodale Verteilungen nützliche monotone Transformation aus der Familie der sogenannten *Power*-Transformationen ist die parametrische Box-Cox-Transformation, welche im Gegensatz zur z-Transformation eine Reduzierung der Schiefe der Verteilung zum Ziel hat (Box und Cox, 1964). Die Verteilung der Werte einer Variablen wird demzufolge durch diese

Transformation symmetrischer und einer Normalverteilung ähnlicher, indem die Transformation Bereiche der Verteilungsfunktion mit niedriger Varianz streckt und solche mit hoher Varianz staucht. Die Transformationsvorschrift für den j-ten Wert der Variablen $x^{(i)}$ lautet

$$\tilde{x}_j^{(i)} = \begin{cases} \frac{\left(x_j^{(i)}\right)^{\lambda} - 1}{\lambda} & \text{falls } \lambda \neq 0 \\ \ln\left(x_j^{(i)}\right) & \text{falls } \lambda = 0 \end{cases} . \tag{3.73}$$

Eine geeignete Wahl für den Transformationsparameter λ ergibt sich über die *Maximum-Likelihood*-Methode (vgl. Kapitel 3.3.2) oder empirisch durch simples Austesten verschiedener Werte. In der Praxis übernehmen geeignete Softwarepakete die Bestimmung von λ.
Allerdings ist die Box-Cox-Transformation nur für positive Werte von $x_j^{(i)}$ definiert. Eine Verallgemeinerung für reellwertige Variablen stellt die Yeo-Johnson-Transformation dar (Yeo und Johnson, 2000). Die Transformationsvorschrift dafür lautet:

$$\tilde{x}_j^{(i)} = \begin{cases} \frac{\left(x_j^{(i)}+1\right)^{\lambda} - 1}{\lambda} & \text{falls } \lambda \neq 0 \wedge x_j^{(i)} \geq 0 \\ \ln\left(x_j^{(i)} + 1\right) & \text{falls } \lambda = 0 \wedge x_j^{(i)} \geq 0 \\ -\frac{\left(-x_j^{(i)}+1\right)^{2-\lambda} - 1}{2-\lambda} & \text{falls } \lambda \neq 2 \wedge x_j^{(i)} < 0 \\ -\ln\left(-x_j^{(i)} + 1\right) & \text{falls } \lambda = 2 \wedge x_j^{(i)} < 0 \end{cases} . \tag{3.74}$$

Eine Kombination von z- und Box-Cox- bzw. Yeo-Johnson-Transformation ist ebenfalls möglich, sodass zentrierte, skalierte und zugleich schiefe-reduzierte Variablenwerte erzeugt werden können.

3.5.2 Resampling zur Balancierung von Datensätzen

Sind die Werte der abhängigen Variablen y in einem Datensatz sehr ungleich verteilt, wirkt sich dies bei der Modellbildung von statistischen Vorhersageverfahren aus. Viele Optimierungsverfahren der Modellbildung schneiden die Modellparameter auf stark vertretene Klasse(n) bzw. Wertebereiche des Prädiktanden zu, während die übrigen kaum Einfluss auf die Schätzung der Modellparameter haben. Ein Beispiel hierfür ist die globale Minimierung einer Kostenfunktion in vielen statistischen Verfahren, welche dann größtenteils die Kosten der überrepräsentierten Klasse(n)/Wertebereiche darstellt (z. B. Gleichungen (3.20) und (3.44)). Vorhersagen von Regressionsverfahren können dadurch lediglich eine geringe Schärfe der Schätzwerte des Prädiktanden aufweisen, d. h. die Vorhersagen variieren nur in einem kleinen Wertebereich und können die Variabilität der beobachteten Werte nicht vollständig abbilden. Studien von Weiss und Provost (2001), Batista et al. (2004) und weiteren legen nahe, dass bestimmte, sogenannte *Resampling*-Methoden (in diesem Abschnitt: RSP-Methoden) die Probleme verringern können, die ein solcher schiefer (unbalancierter) Datensatz mit sich bringt. Jedoch verhalten sich statistische Vorhersageverfahren sehr unterschiedlich nach der Anwendung verschiedener RSP-Methoden auf einen Trainingsdatensatz, sodass Kuhn und Johnson (2013) konstatieren, dass sich keine allgemeingültige Aussage über den Nutzen von solchen Methoden treffen lässt.
Die in der vorliegenden Arbeit betrachteten abhängigen Variablen sind in der verfügbaren Stichprobe sehr ungleich verteilt, wie die Analysen in Kapitel 5 zeigen werden. Zudem deuten die dortigen Untersuchungen darauf hin, dass ein RSP für die Modellstudien in Kapitel 6 vorteilhaft sein könnte. Letztlich ist eine optimale Wahl für eine RSP-Methode zur Balancierung der Klassen bzw. des Wertebereichs des Prädiktanden im Trainingsdatensatz bei der Bildung eines finalen, optimal angepassten Modells zu treffen. Eine Vorstellung verschiedener RSP-Methoden, die getestet wurden, folgt unten.

Für binäre Klassifikationsverfahren sei zuvor das Klassenverhältnis

$$\rho_K = \frac{N_{\mathscr{K}_{klein}}}{N_{\mathscr{K}_{gross}}} \tag{3.75}$$

eingeführt, welches das Verhältnis der Anzahl von Ereignissen, welche die unterrepräsentierte Klasse $\mathscr{K}_{klein}$ darstellen, zu den Nicht-Ereignissen in der überrepräsentierten Klasse $\mathscr{K}_{gross}$ angibt. Der Trennwert der abhängigen Variablen y, der beide Klassen separiert, wird fortan als Klassentrennwert bezeichnet. Eine RSP-Methode kann im Fall von ursprünglich sehr kleinen Werten von ρ_K eine Vergrößerung des Klassenverhältnisses im Vergleich zum originalen Trainingsdatensatz bewirken. Dies bietet speziell für den *Random Forest* die Möglichkeit einer Reduzierung der Anzahl von Entscheidungsbäumen N_{Baum} und somit des Rechenaufwands, da der Entscheidungstrennwert μ aus Gleichung (3.71) deutlich größere Werte annehmen kann.

Um den Wertebereich des Prädiktanden ausgeglichener zu repräsentieren, existieren zwei verschiedene RSP-Methoden, welche beide in den Modellstudien in Kapitel 6 und Anhang B Anwendung finden: das *Undersampling* und das *Oversampling*, welche in diesem Kapitel aufgrund der häufigen Verwendung mit USP und OSP abgekürzt werden. Die prinzipielle Idee hinter dem USP ist eine maßgebliche Reduktion des Anteils der überrepräsentierten Klasse(n)/Werte, während das OSP aus den Datenpunkten der unterrepräsentierten Klasse(n)/Werte zusätzliche fiktive Datenpunkte geschickt generiert (s. u.). Beim USP erfolgt also eine Verkleinerung des Trainingsdatensatzes, während das OSP diesen vergrößert. Darüber hinaus ist es auch möglich, zunächst ein USP und direkt im Anschluss ein OSP durchzuführen.

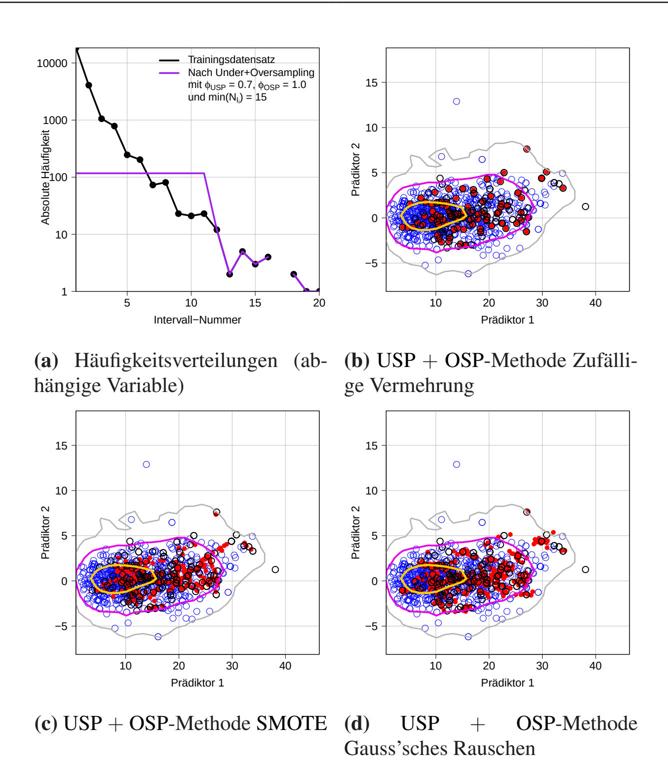
Die im Folgenden beschriebenen Techniken, die jeweils eine Variante des USP und des OSP darstellen, sind generell auf jede beliebige Variable im Datensatz anwendbar, d. h. ein RSP muss nicht zwangsweise bezüglich der abhängigen Variablen y erfolgen. Die späteren Untersuchungen wenden das RSP jedoch ausschließlich bezüglich y an. Alle RSP-Methoden wirken sich durch die Modifikation des Trainingsdatensatzes offensichtlich auch auf die

Bestimmung der Wichtigkeit der Prädiktoren aus. Abbildung 3.3 ergänzt zur Veranschaulichung die nachfolgende Beschreibung von USP und OSP bzw. deren Kombination.

Undersampling

Der Wertebereich der abhängigen Variablen im Trainingsdatensatz wird in $N_{USP} \ll N_{Tr}$ Intervalle I_i aufgeteilt, z. B. $N_{USP} = 20$ (Abbildung 3.3a). Darin steht N_{Tr} führt die Anzahl von Datenpunkten im Trainingsdatensatz. Es folgt die Bestimmung der Anzahl von Datenpunkten in jedem Intervall (N_{I_i}). Anschließend entfernt der Algorithmus aus allen Intervallen, in denen sich mehr Datenpunkte finden als durch ein bestimmtes Perzentil ϕ_{USP} vorgegeben ($N_{\phi_{USP}}$, gerundet), so viele Datenpunkte, dass in jedem Intervall $N_{I_i} \leq N_{\phi_{USP}}$ gilt. ϕ_{USP} wird im Folgenden auch Balanceparameter genannt. Anschaulich gesprochen schneidet das USP einfach den Hügel der Verteilungsfunktion auf ein bestimmtes Niveau ab, das durch ϕ_{USP} kontrolliert wird. Der Balanceparameter ϕ_{USP} fällt in der Regel nicht mit dem Klassentrennwert zusammen. Es ist demnach ebenfalls möglich, dass der Algorithmus bei Klassifikationsverfahren neben Datenpunkten aus $\mathcal{K}_{gross}$ auch solche aus $\mathcal{K}_{klein}$ entfernt.

Durch das USP wächst das Klassenverhältnis ρ_K bei Klassifikationsverfahren an. Je kleiner ϕ_{USP} ist, desto größer ist die Reduzierung der Anzahl von Datenpunkten im Trainingsdatensatz. Bei sehr ungleicher Verteilung sind auch für recht hohe Werte von ϕ_{USP} starke Reduzierungen der Größe des Trainingsdatensatzes möglich. Das USP bewirkt darüber hinaus eine Modifikation der Verteilungsfunktion der den Datenpunkten zugeordneten unabhängigen Variablen im Trainingsdatensatz, welche weiterhin einen großen Wertebereich abdeckt (blaue und schwarze Kreise in den Abbildungen 3.3b−d). Die Verkleinerung des Trainingsdatensatzes hat nicht zwangsläufig eine Verkleinerung der Varianz der Werte der unabhängigen



(a) Häufigkeitsverteilungen (abhängige Variable)

(b) USP + OSP-Methode Zufällige Vermehrung

(c) USP + OSP-Methode SMOTE

(d) USP + OSP-Methode Gauss'sches Rauschen

Abbildung 3.3: (a) Häufigkeitsverteilung einer abhängigen Variablen im originalen Trainingsdatensatz einer exemplarischen Modellstudie ($N_{Tr} = 25\,000$; schwarze Punkte und Linie) sowie im modifizierten Trainingsdatensatz nach der kombinierten Anwendung eines USP ($N_{USP} = 20$, $\phi_{USP} = 0{,}7$) und eines anschließenden OSP ($\phi_{OSP} = 1{,}0$, $N_{I,min} = 15$), welches auf $N_{\phi_{USP}} = 117$ führt (violette Linie). Das USP sortiert dabei Datenpunkte aus den Intervallen I_1 bis I_6 aus, während das OSP für die Intervalle I_7 bis I_{11} fiktive Datenpunkte generiert. Die Reduzierung der Größe des Trainingsdatensatzes liegt bei rund 94,7 %. (b)−(d) Darstellung des RSP im Raum der Prädiktoren. Die Linien kennzeichnen die Häufigkeitsverteilung im originalen Trainingsdatensatz, abgeleitet aus einer 2D-Kerndichteschätzung mit Gaußkern und 30×30 Boxen im Wertebereich der Prädiktoren (grau: 100 ppm; magenta: 1 ‰; orange: 1 %; z. B. Venables und Ripley, 2013). Blaue und schwarze Kreise kennzeichnen Datenpunkte, die nach dem USP übrig bleiben, wobei die schwarzen solche markieren, die in den für das OSP relevanten Intervallen liegen. Rote Punkte sind fiktive Datenpunkte, die die jeweilige OSP-Methode generiert. Die Abbildungen entstammen einer exemplarischen Untersuchung mit der abhängigen Variablen Lebensdauer von konvektiven Zellen und den Prädiktoren DLS (Prädiktor 1; $\mathrm{m\,s^{-1}}$) und LI (Prädiktor 2; K) im Rahmen der Modellstudien, die in Kapitel 6 vorgestellt werden.

Variablen zur Folge. Insgesamt geht jedoch durch das USP im Vergleich zum originalen Trainingsdatensatz Information verloren.

Oversampling

Das OSP verfolgt den umgekehrten Ansatz zum USP und seine numerische Umsetzung erfolgt auf sehr ähnliche Weise. Anschaulich gesprochen hebt das OSP einen Teil des Schwanzes der Verteilungsfunktion auf ein bestimmtes Niveau an, das durch einen Balanceparameter ϕ_{OSP} kontrolliert wird. Die Intervalle, in die weniger als $N_{I,min} = \min(N_{I_i})$ fallen (z. B. 15 Datenpunkte), bleiben jedoch unberührt (Abbildung 3.3a). Die Einführung von $N_{I,min}$ ist darin begründet, dass zu wenige Datenpunkte in einem Intervall die Variabilität der unabhängigen Variablen nicht hinreichend gut repräsentieren. Extrem selten beobachtete Wertebereiche des Prädiktanden erhalten dadurch allerdings noch weniger Gewicht. Die Generierung fiktiver Datenpunkte kann mit unterschiedlichen Methoden erfolgen, von denen drei ausgewählt und getestet werden (s. Anhang B). Allen Methoden gemein ist eine Generierung fiktiver Datenpunkte (rote Punkte) auf der Basis der in den jeweiligen Intervallen vorhandenen Datenpunkte (schwarze Kreise in den Abbildungen 3.3b−d). Dies bedeutet, dass die fiktiven Datenpunkte im statistischen Sinne nicht unabhängig von den vorhandenen sind. Solange die Evaluation der Vorhersageverfahren jedoch mit einem Testdatensatz erfolgt, der keinem RSP unterzogen wurde und die ursprüngliche Verteilung der Werte der Prädiktanden näherungsweise widerspiegelt, ist diese Modifikation der Trainingsdaten eine valide Methode zur Balancierung (Kuhn und Johnson, 2013).

Die erste OSP-Methode vervielfacht Datenpunkte in den betroffenen Intervallen durch zufälliges Ziehen mit Zurücklegen, ähnlich zu einer von Ling und Li (1998) untersuchten Methode (Abbildung 3.3b). Die zweite OSP-Methode basiert auf der von Chawla et al. (2002) eingeführten Methode SMOTE (*Synthetic Minority Oversampling Technique*), welche neue Datenpunkte für jedes betroffene Intervall separat generiert, und zwar auf zufälligen Positionen entlang von Liniensegmenten im Raum der Prädiktoren

zwischen ausgewählten benachbarten Datenpunkten (Abbildung 3.3c). Die dritte OSP-Methode basiert auf der von Lee (1999) vorgestellten Methode und generiert in den betroffenen Intervallen fiktive Datenpunkte im Raum der Prädiktoren zufällig innerhalb eines vorgegebenen Radius um vorhandene Datenpunkte, der auf einem vorgegebenen Anteil (häufig 10 %) an der Standardabweichung der Prädiktoren beruht (Abbildung 3.3d).
Durch das OSP wächst das Klassenverhältnis ρ_K bei Klassifikationsverfahren ebenfalls an. Eine zu große Anzahl von fiktiven Datenpunkten kann allerdings dazu führen, dass häufig redundante Informationen aus denselben vorhandenen Datenpunkten verwendet werden, welche die Modellbildung anschließend zu stark beeinflussen. Ling und Li (1998) merken an, dass OSP-Methoden die Vorhersagen von Verfahren, die auf einer globalen Optimierung beruhen, kaum signifikant verbessern, wie beispielsweise die Methode der kleinsten Quadrate zur Minimierung einer globalen Kostenfunktion (s. o.; vgl. Kapitel 3.3). Verfahren wie z. B. Entscheidungsbäume und somit auch *Random Forests* können von einem OSP profitieren, weil sie den Zustandsraum der Prädiktoren in feinere Unterräume unterteilen können (vgl. Kapitel 3.4; Lee, 1999).

Kombination von *Undersampling* und *Oversampling*

Den drei vorgestellten OSP-Methoden kann auch die Anwendung eines USP vorausgehen (Abbildung 3.3). Dies geschieht in der vorliegenden Arbeit dergestalt, dass nach dem USP ein OSP bezogen auf $N_{\phi_{USP}}$, d. h. mit $\phi_{OSP} = 1{,}0$ nachfolgt. Wie oben beschrieben, stehen hierfür drei verschiedene Methoden zur Verfügung. Die kombinierte Methode schneidet den Hügel der Verteilungsfunktion folglich wie beim reinen USP auf ein bestimmtes Niveau ab und hebt einen Teil des Schwanzes auf dieses Niveau an. Extrem selten beobachtete Wertebereiche des Prädiktanden erhalten wegen der Wahl $N_{I,min} = 15$ dadurch wie beim reinen OSP noch weniger Gewicht. Chawla et al. (2002) zeigten beispielsweise für ein auf einem Entscheidungsbaum basierendes Klassifikationsverfahren, dass eine Kombination von USP und SMOTE die

Vorhersagegüte im Vergleich zur Verwendung des originalen Trainingsdatensatzes oder zu einem reinen USP verbessern kann. Für $\phi_{USP} \rightarrow 0$ ($\phi_{USP} \rightarrow 1$) geht die hier beschriebene Kombination von USP und OSP in ein reines, extremes USP (OSP) über.

3.6 Gütemaße für die Evaluation

Zur Evaluierung von Vorhersagen existiert eine Vielzahl von unterschiedlichen Gütemaßen, welche verschiedene Eigenschaften des Datensatzes und Aspekte der Vorhersagegüte beleuchten können. Oft ist es zudem zur differenzierten Evaluation hilfreich oder gar erforderlich, die Vorhersagen grafisch mit den entsprechenden Beobachtungen zu vergleichen. In der vorliegenden Arbeit sollen sowohl Grafiken als auch bestimmte Maßzahlen einen übersichtlichen Blick auf die Vorhersagegüte ermöglichen (vgl. Kapitel 6.2 bis 6.4). Die Vorhersageverfahren aus den Kapiteln 3.3 und 3.4 untersuchen für konvektive Zellen die Prädiktanden Lebensdauer und maximale Zellfläche anhand von sogenannten Zellobjekten, die auf der Basis von Radardaten bestimmt werden (s. Kapitel 4.1). Dabei produzieren die Verfahren für die Prädiktanden diskrete oder kontinuierliche Vorhersagewerte aus den jeweiligen Testdatensätzen, welche den entsprechenden Beobachtungswerten gegenübergestellt werden. Da die Beobachtungswerte nur ein möglichst realitätsnahes Abbild der tatsächlich aufgetretenen Zellattribute sind, wird im Folgenden weiterhin von der Evaluation und nicht von der Verifikation der Vorhersagen gesprochen.

Die Einführung von Gütemaßen geschieht mit Hilfe von Beispielen der abhängigen Variablen Lebensdauer, ist jedoch gleichermaßen auf die Zellfläche anwendbar. Für die Lebensdauer existieren zwar nur diskrete Werte im Abstand von 5 min, welcher der zeitlichen Auflösung der Informationen aus den

Radarmessungen entspricht. Dennoch kann sie als quasi-kontinuierlich angesehen und bei Regressionsverfahren wie eine kontinuierliche Variable behandelt werden. Für spezielle Aspekte zur Durchführung der Evaluation der in den Kapiteln 6.2 bis 6.4 diskutierten Modellstudien sei auf Kapitel 6.1.2 verwiesen.

3.6.1 Kategorische Evaluation

Die kategorische Evaluation dient dazu, Vorhersagen von Klassifikationsverfahren auszuwerten (Wilks, 2006). Um überhaupt eine Klassifikation vornehmen zu können, muss eine Einteilung kontinuierlicher abhängiger Variablen in verschiedene Kategorien (Klassen) stattfinden. Die *i*-te Klasse wird im Folgenden mit $\mathscr{K}_i$ bezeichnet. Der einfachste binäre Fall, den eine logistische Regression und ein *Random Forest* modellieren können, ist demnach eine Aufteilung in zwei Klassen wie beispielsweise Zellobjekte mit kurzer und langer Lebensdauer. Wie in Kapitel 3.4.2 dargestellt, kann der *Random Forest* auch multikategorische Vorhersagen treffen.
Bei der Konstruktion der zwei Klassen stellt sich nicht nur die Frage nach einem geeigneten Klassentrennwert zwischen diesen (vgl. Kapitel 3.5.2), sondern auch die Frage nach einer sinnvollen und fairen Evaluation der Vorhersagen. Läge beispielsweise der Klassentrennwert für die Lebensdauer bei $\tau = 60\,\mathrm{min}$, kann dann eine Vorhersage einer kurzen Lebensdauer für ein Zellobjekt, das eine Lebensdauer von wenigen Minuten mehr als 60 min aufweist, als falsch bezeichnet werden? Man kann diese Problematik zumindest abfedern, indem man Objekte mit kurzer und langer Lebensdauer klarer separiert. Dies lässt sich beispielsweise durch die Wahl eines symmetrischen Übergangsbereichs $[\tau - \tau' ; \tau + \tau']$ umsetzen, der bei der Evaluation keine Berücksichtigung findet. Je größer τ' ist, desto deutlicher ist die Separation. Eine qualitative und quantitative Untersuchung des Einflusses von τ' sowie der Wahl des Klassentrennwerts τ ist in Anhang B für eine beispielhafte Vorhersage der Lebensdauer dargestellt.

Tabelle 3.1: Kontingenztabelle für die binäre Evaluation nach Heidke (1926).

Beobachtung → Vorhersage ↓	Ereignis (J)	Nicht-Ereignis (N)
Ereignis (J)	*a* Treffer (*Hit*)	*b* Falscher Alarm (*False Alarm*)
Nicht-Ereignis (N)	*c* Versäumnis (*Miss*)	*d* Korrekte Nicht-Vorhersage (*Correct Rejection*)

Ein Spezialfall der kategorischen Evaluation ist im Fall von nur zwei Klassen die binäre Evaluation, welche für viele andere meteorologische Fragestellungen Anwendung findet, insbesondere auch im Bereich der Vorhersage konvektiver Zellen: Tritt ein Gewitter auf? Produziert eine Zelle Hagel? Wird sich heute in einer konvektiven Zelle ein Tornado entwickeln? All diese Fragen können mit ja oder nein beantwortet werden – ein Ereignis tritt also ein oder eben nicht. Analog stellt sich hier nun beispielsweise die Frage: Wird eine detektierte Zelle eine lange Lebensdauer haben? Zur Quantifizierung verschiedener Gütemaße ist für solche Fragestellungen eine Kontingenztabelle hilfreich, welche Vorhersagen und Beobachtungen gegenüberstellt (Tabelle 3.1). Die Buchstaben a, b, c und d stehen darin für die jeweilige Anzahl von registrierten Zellobjekten. Die Summe der vier Werte ergibt den Umfang des zur Evaluation verwendeten Teils des Testdatensatzes $N'_{Te} = a+b+c+d \leq N_{Te}$ mit dem Gesamtumfang des Testdatensatzes N_{Te}. Dabei ist N'_{Te} abhängig von der Wahl des Übergangsbereichs, wobei im Fall ohne Übergangsbereich gilt: $N'_{Te} = N_{Te}$.

Deterministische Gütemaße

Die Trefferrate (*Hit Rate*, H) gibt Auskunft darüber, wie groß der Anteil der korrekten J-Vorhersagen an der Gesamtzahl von J-Beobachtungen ist (Tabelle 3.2; z. B. Doswell et al., 1990). Die Fehlalarmrate (*False Alarm Rate*, F) hingegen zeigt an, wie groß der Anteil der falschen J-Vorhersagen an der Gesamtzahl von N-Beobachtungen ist. H soll folglich möglichst groß und F möglichst klein sein. Anzustreben ist eine klar positive Differenz aus beiden. Diese wird auch als *True Skill Statistic* (TSS) oder Peirce Skill Score bezeichnet (Peirce, 1884).
Das Erfolgsverhältnis (*Success Ratio*, *SR*) und das Fehlalarmverhältnis (*False Alarm Ratio*, *FAR*) geben die Anteile der korrekten bzw. falschen J-Vorhersagen an allen J-Vorhersagen an (z. B. Doswell et al., 1990). Eines der beiden Maße genügt bereits für eine Quantifizierung der Belastbarkeit einer J-Vorhersage. Ist das *FAR* beispielsweise sehr groß, so tritt in den meisten Fällen einer J-Vorhersage ein Nicht-Ereignis auf.
Die Genauigkeit einer Vorhersage ist intuitiv durch den Anteil aller korrekten Vorhersagen an allen Vorhersagen gegeben, welcher als *Proportion Correct* (*PC*) bezeichnet wird (Finley, 1884). Der *Critical Success Index* (*CSI*) vernachlässigt korrekte N-Vorhersagen bei der Bestimmung der Genauigkeit (Gilbert, 1884).
Der Bias (B) ist ein Indikator für sogenanntes *Over*- und *Underforecasting*. Er vergleicht die Anzahl von J-Vorhersagen mit der Anzahl von J-Beobachtungen. Ist beispielsweise $B \gg 1$, so sagt ein Modell oft Ereignisse vorher, jedoch ohne dass das Ereignis in der Folge eintritt (*Overforecasting*).
Das *Odds Ratio* (*OR*) gibt das Chancenverhältnis für die Chance an, ein Ereignis zu beobachten (z. B. Stephenson, 2000). Ist $OR > 1$, so ist die Chance für eine J-Beobachtung bei einer J-Vorhersage größer als bei einer N-Vorhersage.

Tabelle 3.2: Übersicht über verschiedene (deterministische) Gütemaße im Zusammenhang zur Kontingenztabelle (Tabelle 3.1), sowie deren Berechnung, Wertebereich $\mathbb{W}$ und den Aspekt der Vorhersage, den sie beleuchten. Der optimale Wert für jedes Gütemaß ist in fetter Schrift bzw. in runden Klammern dargestellt. Die Abkürzungen für die Aspekte lauten U (Unterscheidungsvermögen), B (Belastbarkeit), G (Genauigkeit), OU (*Over-* oder *Underforecasting*) und C (Chancenverhältnis). Der Index Z steht hier für Zufall. Die Formelzeichen in der Berechnung des *ETS* und *HSS* stehen für $\mathcal{N}_1 = (N'_{Te} - a)(N'_{Te} - d) - bc$ und $\mathcal{N}_2 = (N'_{Te} - a)(N'_{Te} - d) - 2bc + ad$.

Gütemaß (Score)	Berechnung	$\mathbb{W}$	Aspekt
Hit Rate	$H = \frac{a}{a+c}$	$[0; \mathbf{1}]$	U
(Prob. of Detection)			
False Alarm Rate	$F = \frac{b}{b+d}$	$[\mathbf{0}; 1]$	U
(Prob. of False Det.)			
False Alarm Ratio	$FAR = \frac{b}{a+b}$	$[\mathbf{0}; 1]$	B
Succes Ratio	$SR = \frac{a}{a+b}$	$[0; \mathbf{1}]$	B
Proportion Correct	$PC = \frac{a+d}{N'_{Te}}$	$[0; \mathbf{1}]$	G
Critical Success Index	$CSI = \frac{a}{a+b+c}$	$[0; \mathbf{1}]$	G
(Threat Score)			
Bias	$B = \frac{a+b}{a+c}$	$\mathbb{R}_0^+$ $(\mathbf{1})$	OU
Odds Ratio	$OR = \frac{ad}{bc}$	$\mathbb{R}_0^+$ (∞)	C
(Chancenverhältnis)			
Gütemaß (Skill Score)			
True Skill Statistic	$TSS = H - F$	$[-1; \mathbf{1}]$	U
(Peirce Skill Score)			
Equitable Threat Score	$ETS = \frac{ad-bc}{\mathcal{N}_1}$	$[-\frac{1}{3}; \mathbf{1}]$	*CSI* vs. CSI_Z
(Gilbert Skill Score)			
Heidke Skill Score	$HSS = \frac{2(ad-bc)}{\mathcal{N}_2}$	$[-1; \mathbf{1}]$	*PC* vs. PC_Z

Der *Equitable Threat Score* (ETS) nach Gilbert (1884) sowie der Heidke Skill Score (HSS) nach Heidke (1926) vergleichen auf Basis des CSI bzw. des PC die Vorhersagen mit einer zufälligen Vorhersage, bei der die Vorhersagen statistisch unabhängig von jeglichen Beobachtungen sind. Ist $ETS > 0$ bzw. $HSS > 0$, so ist die Vorhersage besser als eine solche zufällige Vorhersage.

Das ROC-Diagramm (*Relative/Receiver Operating Characteristic*) kombiniert grafisch H und F (Abbildung 3.4; z. B. Mason, 1982). Hier ist ein Vergleich von verschiedenen Realisierungen eines Vorhersagemodells möglich, wobei Realisierung ein bestimmtes Modell-Setup bezeichnet (s. Kapitel 6.2.2). In den Kapiteln 6.2 bis 6.4 dient das ROC-Diagramm dazu, einen Überblick über Realisierungen der logistischen Regression und des *Random Forests* zu schaffen, die auf verschiedenen Werten des Entscheidungstrennwerts μ in Bezug auf $\hat{p}(y = 1 \mid \mathbf{x} = \mathbf{x}_j)$ bei der Zuordnung der Vorhersagen zu den beiden Klassen des Prädiktanden beruhen (vgl. Gleichungen (3.36), (3.71) und (3.72)).

Eine perfekte Modellvorhersage liegt im Diagramm links oben, sodass der Abstand $D = 2^{-0,5} TSS$ zur Diagonalen maximal, nämlich $D = 2^{-0,5}$ ist. Liegen alle Realisierungen eines Modells links oben, ist die Fläche unter der ROC-Kurve, die *Area Under the Curve* (AUC), maximal, nämlich $AUC = 1$. Ist $AUC \leq 0,5$, so befindet sich die ROC-Kurve (größtenteils) rechts der Diagonalen, was auf eine schlechtere Vorhersagegüte als bei einer unabhängigen zufälligen Vorhersage hindeutet.

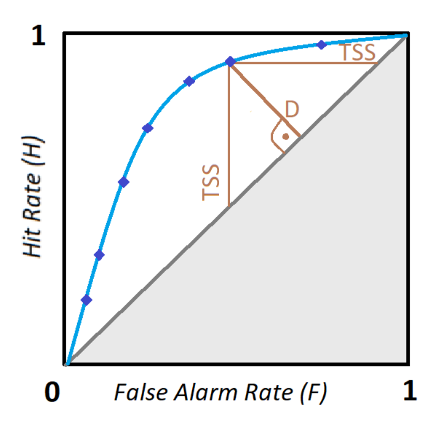


Abbildung 3.4: Schematische Darstellung eines ROC-Diagramms. Die dunkelblauen Punkte stellen die Werte von H und F für unterschiedliche Realisierungen einer Vorhersage dar und sind zur ROC-Kurve verbunden. In braun ist der Bezug zwischen der TSS und dem Abstand der ROC-Kurve zur Diagonalen geometrisch verdeutlicht.

Probabilistische Gütemaße

Während die ROC-Kurve zur Bestimmung des (je nach Fragestellung) optimalen Entscheidungstrennwerts μ anhand verschiedener Realisierungen der Modelle mit unterschiedlichen Werten für μ nützlich ist, dient eine mehrfache Modellbildung mit unterschiedlichen Trainings- und/oder Testdatensätzen dazu, die Abhängigkeit von den verwendeten Trainings- und/oder Testdatensätzen zu quantifizieren (s. Kapitel 6.1.1; z. B. James et al., 2013).
Zum einen ist ein Vergleich mehrerer ROC-Kurven möglich, die auf verschiedenen Aufteilungen des Datensatzes in Trainings- und Testdaten beruhen. Finden sich große qualitative Unterschiede in den Verläufen der ROC-Kurven unter Verwendung unterschiedlicher Trainingsdaten bei immer den gleichen

Testdaten, so ist das Modell stark von der Auswahl der Trainingsdaten abhängig, d. h. die Trainingsdaten sind nicht repräsentativ für den gesamten Datensatz. Finden sich große qualitative Unterschiede in den Verläufen der ROC-Kurven unter Verwendung der gleichen Trainingsdaten und unterschiedlicher Testdaten, so sind die Testdaten nicht repräsentativ für den gesamten Datensatz. In beiden Fällen fällt es schwer, ein optimales μ zu bestimmen.
Zum anderen kann man für eine bestimmte Realisierung eines Modells alleine, also mit festem μ, die Sensitivität bezüglich der Trainings- und Testdaten untersuchen. Eine Gruppe von Vorhersageläufen eines Modells mit unterschiedlichen Ausgangsdaten wird als Ensemble bzw. Modellensemble bezeichnet. Die Gruppe setzt sich aus sogenannten Ensemblemitgliedern zusammen. Der Ensembleansatz hat die Quantifizierung der Unsicherheit (Schwankungsbreite) der Vorhersagen zum Ziel. Die Varianz der Gütemaße innerhalb des Ensembles ist wiederum ein Maß für die Robustheit des Modells bezüglich des Datensatzes. Die Verteilung der Vorhersagen der Ensemblemitglieder für einzelne Zellobjekte bietet die Möglichkeit einer probabilistischen Vorhersage der jeweiligen Lebensdauer. Je höher die Varianz der Vorhersagen der einzelnen Mitglieder ist, desto unsicherer ist die Vorhersage (s. u.).

Der Ensembleansatz ermöglicht eine zellspezifische Betrachtungsweise bei der Evaluation der Vorhersagen der Ensemblemitglieder. Für jedes der Zellobjekte im Testdatensatz können separat die Vorhersagen der Mitglieder $\hat{y}_j^{(q)} \in \{0;1\}$ in eine gemeinsame Vorhersage für die Wahrscheinlichkeit eines Ereignisses überführt werden, sodass die Ensemblevorhersage für das j-te Zellobjekt lautet:

$$\hat{y}_j^{(ens)} = \frac{1}{N_{ens}} \sum_{q=1}^{N_{ens}} \hat{y}_j^{(q)} \ . \tag{3.76}$$

Zur Beurteilung der Qualität der Vorhersagen bezüglich aller Zellobjekte des Testdatensatzes ist anschließend der Brier Score (*BS*) nützlich. Dieser entspricht im Wesentlichen dem mittleren quadratischen Fehler *MSE* aus Gleichung (3.23), wobei der beobachtete Wert des Prädiktanden y_j nur die Werte 0 (Nicht-Ereignis) und 1 (Ereignis) annimmt (z. B. Wilks, 2006):

$$BS = \frac{1}{N'_{Te}} \sum_{j=1}^{N'_{Te}} \left(y_j - \hat{y}_j^{(ens)} \right)^2 . \tag{3.77}$$

Der ursprünglich von Brier (1950) eingeführte Score berücksichtigte zusätzlich die jeweiligen quadratischen Abweichungen vom Gegenereignis der Beobachtung:

$$BS_{orig} = \frac{1}{N'_{Te}} \sum_{j=1}^{N'_{Te}} \left\{ \left[y_j - \hat{y}_j^{(ens)} \right]^2 + \left[(1 - y_j) - \left(1 - \hat{y}_j^{(ens)} \right) \right]^2 \right\} = 2\, BS . \tag{3.78}$$

Wie für den *MSE* ist der optimale Wert $BS_{opt} = 0$, nach oben bildet $BS = 1$ die Grenze. Eine Vorhersage mit $\hat{y}_j^{(ens)} = 0{,}5 \;\forall\; j$ entspricht der Vorhersage, dass nicht klar ist, ob $y = 0$ oder $y = 1$ eintreten wird (50 %-Vorhersage bzw. unsichere Vorhersage). Dies führt zu $BS = 0{,}25$. Eine Vorhersage, bei der $\hat{y}_j^{(ens)}$ zufällig aus einer uniformen Verteilung der Eintrittswahrscheinlichkeiten gezogen wird, führt zu $BS = 0{,}33$ (zufällige Vorhersage).
Zum Vergleich des *BS* einer Vorhersage mit einer Referenzvorhersage, wie beispielsweise der unsicheren oder zufälligen Vorhersage, ist der Brier Skill Score (*BSS*) geeignet, der über

$$BSS = \frac{BS - BS_{ref}}{BS_{opt} - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{3.79}$$

definiert ist. Für $BSS > 0$ ist die Vorhersagegüte höher als bei der entsprechenden Referenzvorhersage.

Die über alle Zellobjekte gemittelte Standardabweichung der Vorhersagen der Ensemblemitglieder $\hat{y}_j^{(q)}$ stellt ein Maß für die mittlere Schwankungsbreite (*Spread*) des Ensembles gemäß

$$\hat{\bar{\sigma}}_{ens} = \frac{1}{N'_{Te}} \sum_{j=1}^{N'_{Te}} \hat{\sigma}_j^{(ens)} \tag{3.80}$$

dar, mit

$$\hat{\sigma}_j^{(ens)} = \sqrt{\frac{1}{N_{ens}-1} \sum_{q=1}^{N_{ens}} \left(\hat{y}_j^{(q)} - \hat{y}_j^{(ens)} \right)^2}\,, \tag{3.81}$$

welche – gleicher Testdatensatz für jedes Ensemblemitglied vorausgesetzt – halbbinomial ist: Da $\hat{y}_j^{(q)}$ eine Bernoulli-Variable ist, kann $\hat{\sigma}_j^{(ens)}$ nur $\lfloor 0.5\ N_{ens} \rfloor + 1$ Werte annehmen. Je größer $\hat{\bar{\sigma}}_{ens}$ ist, desto deutlicher unterscheiden sich die Vorhersagen der Ensemblemitglieder im Mittel über alle Zellobjekte.

3.6.2 Kontinuierliche Evaluation

Regressionsverfahren haben zum einen den Vorteil, dass die erstellten Vorhersagen einen kontinuierlichen Wertebereich abdecken. Demzufolge erfolgt beispielsweise die Schätzung der Lebensdauer eines Zellobjekts in Minuten. Aus der *Nowcasting*-Perspektive scheint eine solche Schätzung prinzipiell erstrebenswerter als eine Vorhersage, die nur unterscheidet, ob eine konvektive Zelle eine kurze oder lange Lebensdauer haben wird. Details hierzu finden sich in der Diskussion der Ergebnisse der Modellstudien in den Kapiteln 6 und 7. Zum anderen entfällt bei Regressionsverfahren die Einführung eines Klassentrennwerts mit dem zugehörigen Übergangsbereich – zwei Freiheitsgrade, welche die Evaluation der Vorhersagen potentiell weniger klar und eindeutig gestalten.

Die Untersuchung und Evaluation von Regressionsverfahren geschieht in den Kapiteln 6.3.2 und 6.4.2 ausschließlich mit einem Ensemble von Modellen, welche auf unterschiedlichen Trainingsdaten basieren. Von den einzelnen Vorhersagen der Mitglieder $\hat{y}_j^{(q)} \in \mathbb{R}$ oder vom Ensemblemittel $\hat{y}_j^{(ens)}$ ausgehend, welches sich analog zu Gleichung (3.76) berechnen lässt, kann dann beispielsweise der *MSE* bzw. *RMSE* zur Beurteilung der Vorhersagegüte oder über Gleichung (3.80) die mittlere Schwankungsbreite des Ensembles $\hat{\bar{\sigma}}_{ens}$ berechnet werden. Zur grafischen Evaluation ist es hilfreich, die Häufigkeiten der vorhergesagten Werte des Prädiktanden den jeweiligen beobachteten Werten gegenüberzustellen (s. Kapitel 6.1.2).

4 Datengrundlage und Methoden der Datenaufbereitung

Zur Analyse der Zusammenhänge zwischen dem Lebenszyklus konvektiver Zellen und den vorherrschenden atmosphärischen Umgebungsbedingungen erfolgt eine Kombination zweier Datensätze. Der Lebenszyklus konvektiver Zellen über Deutschland wird mit Hilfe von Daten aus dem KONRAD-Verfahren des DWD abgebildet, einem Verfahren zur Zelldetektion und -verfolgung basierend auf Radardaten (Kapitel 4.1). Die atmosphärischen Umgebungsbedingungen werden anhand von Modelldaten des ehemals operationellen NWV-Modells COSMO-EU des DWD untersucht (Kapitel 4.2). Beide Datensätze sind für die vorliegende Arbeit für den Zeitraum der Sommerhalbjahre 2011 – 2016 verfügbar (1. April – 30. September), welche den Zeitraum eines Jahres darstellen, innerhalb dessen die meisten Gewitter in Deutschland auftreten (z. B. Wapler und James, 2015). Diese sechs Sommerhalbjahre werden fortan als Untersuchungszeitraum bezeichnet. Kapitel 4.3 beschreibt die Datenaufbereitung beider Datensätze, wodurch ein geeigneter kombinierter Datensatz für die weiteren Untersuchungen generiert wird.

4.1 Daten aus dem radarbasierten Verfahren KONRAD

4.1.1 Radarmessungen des Deutschen Wetterdienstes

Niederschlagsradare ermöglichen die Fernerkundung von Niederschlag über ein indirektes Messprinzip (z. B. Sauvageot, 1992). Dabei ist der Begriff Radar ein Akronym für *Radio Detection and Ranging*. Ein in den meisten Fällen verwendetes gepulstes Radar sendet kurze Pulse gebündelter elektromagnetischer Strahlung im Frequenzbereich von Radiowellen (genauer: Mikrowellen) in eine gewünschte Raumrichtung, die von verschiedenen Streukörpern in der Atmosphäre wie beispielsweise Hydrometeoren im durchlaufenen Luftvolumen gestreut werden. Der rückgestreute Anteil wird von der Radarantenne empfangen und in ein digitales Signal umgewandelt. Aus den Eigenschaften des empfangenen Signals (Radarechos) können unter anderem Rückschlüsse auf die Position von Streukörpern und deren Rückstreuung gezogen werden. Die (Radar-)Reflektivität beschreibt die Summe aller Rückstreuquerschnitte der in einem Teil des Strahlvolumens vorhandenen Streukörper. Unter der Annahme von reiner Rayleigh-Streuung ist der Rückstreuquerschnitt eines Streukörpers proportional zur sechsten Potenz seines Durchmessers. Meteorologisch relevant ist schließlich der (Radar-)Reflektivitätsfaktor z, der das sechste statistische Moment des Streukörperdurchmessers D mit einer bestimmten Anzahldichteverteilung $n(D)$ darstellt:

$$z = \int_{D_{min}}^{D_{max}} n(D) D^6 \mathrm{d}D \; . \tag{4.1}$$

Für diesen gilt folgender Zusammenhang zu der vom Radar empfangenen Leistung P_e:

$$z = c P_e \frac{d^2}{|K|^2} \; . \tag{4.2}$$

Darin kennzeichnet c die sogenannte Radarkonstante, die sämtliche technischen Faktoren des Radargeräts zusammenfasst, d die Entfernung der Streukörper und K den komplexen Dielektrizitätsfaktor. Dabei ist für flüssige Wasserpartikel $|K|^2 \approx 0{,}93$ und für Eispartikel etwa $|K|^2 \in (0{,}16\,;0{,}21)$. Die Rückstreuung von einem Eispartikel eines bestimmten Durchmessers ist demnach etwa fünfmal schwächer als die an einem Wassertropfen gleichen Durchmessers. Zur Bestimmung von z aus Radarmessungen wird im operationellen Betrieb des DWD der Dielektrizitätsfaktor von flüssigem Wasser verwendet, was eine Unterschätzung von z im Fall (komplett) gefrorener Hydrometeore zur Folge hat. Da z in seiner ursprünglichen Einheit ($\mathrm{mm}^6\,\mathrm{m}^{-3}$) über viele Größenordnungen variiert, wurde der Reflektivitätsfaktor Z eingeführt:

$$Z = 10\,\log_{10}\left(\frac{z}{\mathrm{mm}^6\,\mathrm{m}^{-3}}\right) \,. \tag{4.3}$$

Die Einheit von Z ist dabei dBZ, wobei Z für Regentropfen typischerweise zwischen 0 (kaum messbarer Niederschlag) und 60 dBZ (heftiger Starkregen) variiert. Für Z kann ein approximativer Zusammenhang zur Regenrate R ($\mathrm{mm\,h}^{-1}$) unter der Annahme einer Anzahldichteverteilung von Regentropfen hergestellt werden, die mit wachsendem Durchmesser exponentiell abnimmt (Marshall und Palmer, 1948). Dieser Zusammenhang ist als Z-R-Beziehung bekannt:

$$Z = aR^b \,. \tag{4.4}$$

Dabei wurden in einer großen Anzahl von Studien empirisch viele verschiedene Werte für die Parameter a und b je nach der Niederschlagsart, der Wolkenstruktur und der Stärke des Niederschlags bestimmt. Der DWD verwendet

für seine C-Band-Radare (Frequenzbereich um 5,6 GHz) bei manchen Radarprodukten eine standardmäßige *Z*-*R*-Beziehung mit den festen Parameterwerten $a = 256$ und $b = 1{,}42$, bei einigen Produkten findet eine mit Hilfe des tatsächlichen Werts des Reflektivitätsfaktors sowie seines horizontalen Gradienten verfeinerte *Z*-*R*-Beziehung Anwendung (Bartels et al., 2004; Weigl, 2015).
Neben dem über P_e bestimmbaren Reflektivitätsfaktor messen Radargeräte mit Doppler-Technik zusätzlich die Radialgeschwindigkeit der Hydrometeore anhand der Frequenzverschiebung der elektromagnetischen Wellen, die auf den Doppler-Effekt zurückzuführen ist (z. B. Blahak, 2005). Polarisationsradare senden und empfangen Wellen auf zwei unterschiedlichen Polarisationsebenen (*Dual-Pol*), welche je nach dem Achsenverhältnis eines Streukörpers unterschiedlich stark zurückgestreut werden (Vivekanandan et al., 1999). Damit können näherungsweise verschiedene Arten von Hydrometeoren klassifiziert werden.

Der Radarverbund des DWD besteht aktuell (Stand: 12. Januar 2021) aus 17 operationellen Niederschlagsradaren mit *Dual-Pol* Doppler-Technik (Abbildung 4.1). Jedes der Einzelradargeräte ist auf einem 20 – 75 m hohen Stahl- oder Betonturm montiert, welcher sich im orografisch gegliederten Gelände Süd- und Mitteldeutschlands auf einem Berggipfel oder in exponiertem Terrain befindet. Dadurch wird eine Abschattung der Radarstrahlen durch Objekte (Gebäude, Vegetation etc.) weitestgehend vermieden. Die Reichweite der Einzelradargeräte beträgt rund 150 km. Die Zusammensetzung des Radarverbunds änderte sich im Untersuchungszeitraum 2011 – 2016 mehrmals (z. B. Winterrath et al., 2017). Im Zuge der Umstellung von *Single*- auf *Dual-Pol*-Technik wurden die Standorte von einigen der zunächst 16 operationellen Einzelradare aus innerstädtischen Bereichen auf nahe gelegene, weniger bebaute Standorte verlegt. Am 3. April 2013 ging am Standort Memmingen das siebzehnte Radar in Betrieb. Während der Technikumstellung

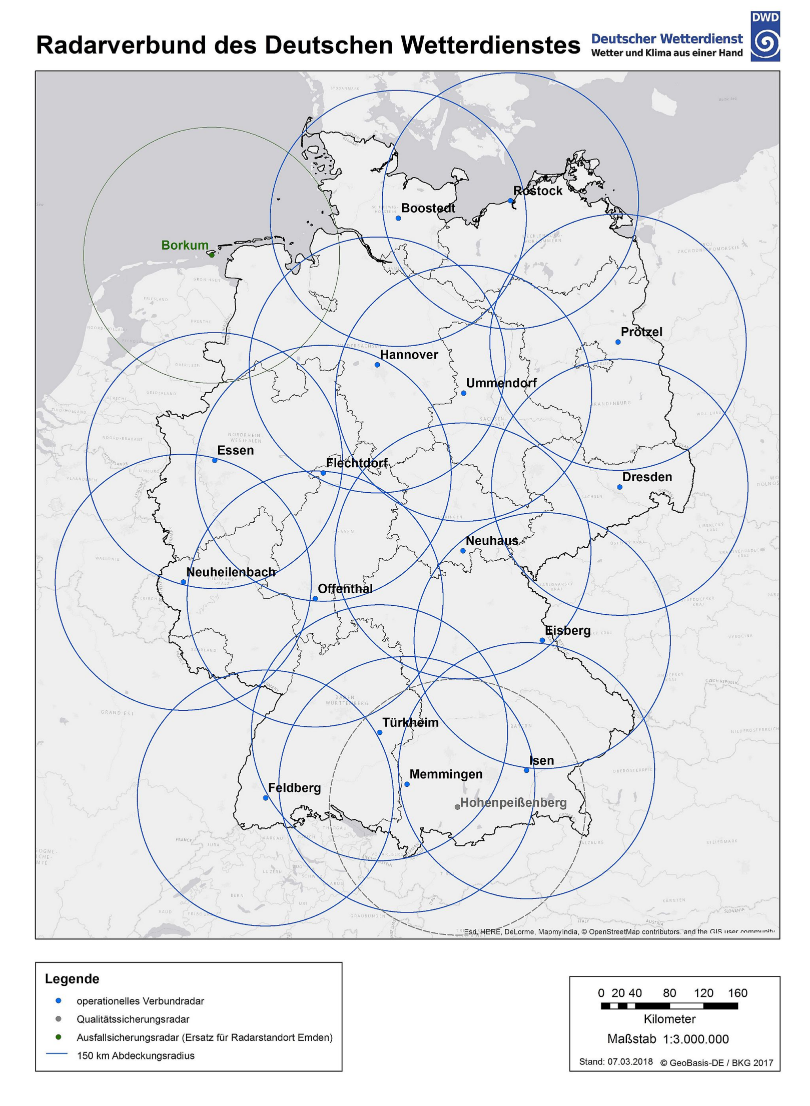


Abbildung 4.1: Karte der Standorte der zum Radarverbund des DWD gehörigen Niederschlagsradare (Stand 07. März 2018; DWD, 2021b).

betrieb der DWD ein mobiles Ausfallsicherungsradar anstelle der operationellen Radare unmittelbar neben den Standorten im Randbereich des Radarverbunds, wo größere Gebiete von keinem der übrigen Radare erfasst werden (Essen, Rostock, Dresden, Feldberg im Schwarzwald). Das Radar Neuheilenbach in der Eifel wurde durch das belgische Radar Wideumont und das Radar Eisberg nahe Moosbach in der Oberpfalz durch das tschechische Radar Brdy abgesichert. Bis zum Ende des Untersuchungszeitraums waren bis auf den Standort Emden, der am 27. Februar 2018 durch den Standort Borkum ersetzt wurde, alle Radare auf *Dual-Pol* Doppler-Technik umgestellt.
Infolge der Technikumstellung sowie gelegentlicher kurzer Datenausfälle (z. B. aufgrund von Wartungsarbeiten) sind die radarbasierten Niederschlagsmessungen über Deutschland keineswegs homogen. Dank einer hochwertigen Filterung von Störeffekten, welche bei Radarmessungen auftreten, und einer Qualitätskontrolle der Radardaten seitens des DWD (z. B. Seltmann und Riedl, 1999) wird die Qualität der Daten deutlich gesteigert, auch wenn einige Effekte nicht korrigiert werden können wie z. B. die räumliche Strahlaufweitung und -abschattung. Die aus den Daten abgeleiteten Radar(bild)produkte sind dennoch für die Anwendung einer automatischen Interpretation konvektiver Zellen geeignet (s. Kapitel 4.1.2). Die Messungen der Einzelradargeräte werden auf ein äquidistantes Gitter projiziert, welches nahezu ganz Deutschland und die Randgebiete verschiedener Nachbarstaaten abdeckt (DWD, 2020). Solche Zusammenstellungen bezeichnet der DWD als Deutschland-Komposits. Ein wichtiges Radarprodukt ist das RX-Produkt, ein Deutschland-Komposit des Reflektivitätsfaktors Z aus dem oberflächennahen und geländefolgenden Niederschlagsscan (Elevationswinkel zwischen 0,5 und 1,8°) mit einer horizontalen Gitterauflösung von $1 \times 1\,\mathrm{km}^2$ und einer zeitlichen Auflösung von 5 min.

4.1.2 Der Zellverfolgungsalgorithmus KONRAD

Die Einführung von KONRAD als operationelles Verfahren im Jahr 2000 zielte darauf ab, das konvektive Unwettergeschehen aus den Radarbildern herauszufiltern und grafisch darzustellen, um Vorhersager*innen und externen Nutzer*innen eine schnelle Übersicht über die von Unwettern gefährdeten Gebiete zu geben und kurzfristig zu treffende Entscheidungen zu erleichtern (Lang et al., 2003). Das Verfahren stellt darüber hinaus ein wichtiges Werkzeug für den automatisierten Warnprozess AutoWARN im *Nowcasting*-Verfahren NowCastMix des DWD dar (vgl. Kapitel 1; James et al., 2018). Außerdem wird es im Rahmen des Systems webKONRAD als Teil des Feuerwehrinformationssystems FeWIS intensiv von Einrichtungen des Katastrophenschutzes genutzt (DWD, 2021d).
Das RX-Produkt des DWD, das für jede Gitterzelle den maximalen Wert von Z aus dem Niederschlagsscan der entsprechenden Einzelradare enthält, stellte während des Untersuchungszeitraums die Grundlage für das automatische Verfahren KONRAD (Konvektionsentwicklung in Radarprodukten) zur Verfolgung von konvektiven Zellen dar[1] (Lang, 2001). KONRAD isoliert aus dem RX-Produkt die Radarechos konvektiver Zellen. Es zeigt ihre aktuelle Position an und zeichnet ihre Zugbahnen auf (s. u.). Darüber hinaus erfolgt die Bestimmung von Warnindikatoren (Warn-*Flags*) bezüglich typischer Begleiterscheinungen (starker) konvektiver Zellen (Hagel, Starkregen und konvektive Starkwindböen) sowie eine Abschätzung der Intensität und des Entwicklungsstadiums der Zellen. Auch extrapoliert KONRAD die Zugbahnen für die nächsten 30 bzw. 60 min auf der Basis der Zellverlagerung während der vorherigen Zeitschritte. Verschiedene Zellattribute wie z. B. die Position,

[1] Bei der operationellen Einführung von KONRAD wurde ursprünglich das PX-Produkt verwendet, welches Z für die jeweiligen Einzelradare in sechs diskreten Klassen (Schwellenwerte: 19, 28, 37, 46, 55 dBZ) beinhaltete.

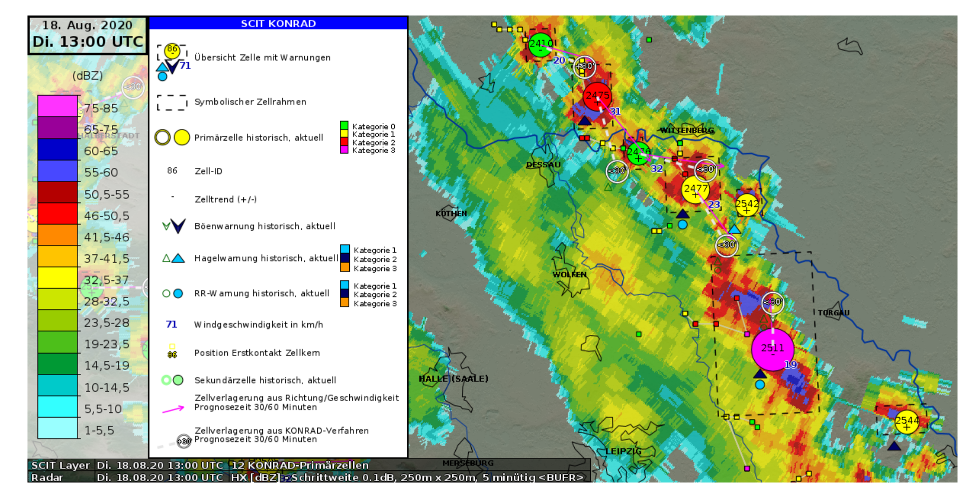


Abbildung 4.2: Visualisierung von Daten aus dem Zellverfolgungsalgorithmus KONRAD, wie sie für Vorhersager*innen des DWD an ihrem Arbeitsplatz zur Verfügung steht. Farbige flächenhafte Felder repräsentieren den Radarreflektivitätsfaktor Z (dBZ), hier aus dem hochaufgelösten HX-Produkt ($250 \times 250\,\mathrm{m}^2$), auf einer geografischen Hintergrundkarte (Ausschnitt nahe Leipzig und Halle an der Saale für den 18. August 2020, 13:00 UTC). Eingefärbte Kreise zeigen die aktuelle Position von KONRAD-Zentroiden, deren Größe mit der Fläche der Zellobjekte skaliert. Die Farbgebung entspricht einer internen Kategorisierung der Intensität. Gezeigt sind nur Primärzellen mit ihrer Identifikationsnummer. Der rechteckige Rahmen, der das jeweilige Zellobjekt einschließt, ist schwarz gestrichelt dargestellt. Historische Positionen der Zentroide sind durch kleine farbige Rechtecke den jeweiligen Intensitätskategorien entsprechend eingezeichnet und durch graue durchgezogene Linien miteinander verbunden. Historische Zentroide können zusätzlich durch ungefüllte Kreise um die Rechtecke herum eingezeichnet werden. Die von KONRAD intern abgeschätzte zukünftige Verlagerungsrichtung ist durch grau gestrichelte Linien visualisiert, magentafarbene Pfeile zeigen die Verlagerungsrichtung aus einer weiteren Methode. An der linken unteren Ecke der mit den Zentroiden assoziierten Kreise sind jeweils die Warnindikatoren in kategorischer Abstufung eingefärbt eingetragen. Auch die Historie der Warnindikatoren kann visualisiert werden. Persönliche Kommunikation von R. Feger und T. Böhme (DWD).

die Zellgröße (horizontale Ausdehnung) und die Verlagerungsrichtung der Zelle werden abgespeichert und in der operationellen Anwendung grafisch aufbereitet und ausgegeben (Abbildung 4.2).

Kurzbeschreibung der Zelldetektion und der Zellverfolgung in KONRAD

KONRAD definiert sogenannte Primärzellen, welche eine zusammenhängende Fläche von Radarbildpixeln darstellen, die einen Schwellenwert von $Z = 46\,\mathrm{dBZ}$ erreichen bzw. überschreiten. Dieser Wert des Reflektivitätsfaktors entspricht in etwa einer Regenrate von $R \approx 23\,\mathrm{mm\,h^{-1}}$. Es müssen mindestens 15 benachbarte Pixel mit einer Überschreitung dieses Schwellenwerts existieren, damit KONRAD eine Zelle detektiert und als Primärzelle abspeichert. Aufgrund dieser Bedingung kann KONRAD als radarbasiertes Verfahren schwache konvektive Zellen nicht detektieren, die den Schwellenwert von Z für weniger als 15 benachbarte Pixel überschreiten. Zudem kann KONRAD auch bei größeren Zellen Teile des Cumulus- und des Dissipationsstadiums konvektiver Zellen nicht gut erfassen. Auswerteverfahren, die auf Satellitendaten basieren, sind hier weniger Grenzen gesetzt. Diese können konvektive Zellen deutlich früher und länger detektieren (vgl. Kapitel 2.4; z. B. Zinner et al., 2008). Aufgrund der erwähnten Kriterien repräsentieren Primärzellen eher einen ausgeprägten Niederschlagsschwerpunkt konvektiver Zellen mit hoher Intensität. Das Umfeld der Zellen mit geringeren Werten des Reflektivitätsfaktors bleibt außer Acht, sodass die Information über die Ausdehnung des gesamten Niederschlagsbereichs der Zellen verloren geht.

Bei der ersten Detektion einer Primärzelle wird ihr aufgrund der zeitlichen Auflösung der Radarbilder von 5 min ein Alter von 2 min zugewiesen. Jeder Primärzelle ordnet KONRAD einen reflektivitätsgewichteten Mittelpunkt (ein sogenanntes Zentroid) sowie einen rechteckigen Rahmen zu, der die Primärzelle so komplett einschließt, dass seine Fläche minimal ist. Die Verfolgung konvektiver Zellen verläuft dergestalt, dass KONRAD für jedes

aktuell registrierte Zentroid im Vorprodukt (5 min vorher) in einem geeigneten Radius ein korrespondierendes Zentroid sucht. Den Radius wählt KONRAD entsprechend einer maximalen realistischen Verlagerungsgeschwindigkeit von $c_{Z,max} = 110\,\mathrm{km\,h^{-1}}$ plus einer möglichen Verlagerung des Zentroids innerhalb der Primärzelle. Dabei sind folgende Rahmenbedingungen relevant: Das aktuelle Zentroid soll nahe am zuvor ermittelten Prognosepunkt liegen, die Strecke soll möglichst kurz sein und die flächenhafte Ausdehnung der Primärzelle soll konsistent sein, d. h. die Fläche von Primärzellen darf sich nicht unrealistisch stark vergrößern oder verkleinern. Ordnet KONRAD einer aktuellen Primärzelle ein korrespondierendes Zentroid im Vorprodukt zu, wird die mit diesem Zentroid assoziierte Primärzelle mit dem Alter t als Vorgängerin der aktuellen Primärzelle angesehen, welche nun ein Alter von $t + 5$ min hat. Primärzellen, für die KONRAD kein korrespondierendes Zentroid im Vorprodukt finden konnte, werden als Neubildung behandelt. Solche, die KONRAD im aktuellen Produkt nicht wiederfindet, werden als zerfallende Zellen behandelt. Zur frühzeitigen Erkennung potentiell neu gebildeter Zellen und zur Erfassung dissipierender Zellen detektiert KONRAD auch Sekundärzellen, die durch die Überschreitung des Schwellenwerts $Z = 37\,\mathrm{dBZ}$ gekennzeichnet sind. Diese werden ebenfalls abgespeichert und wahlweise in der operationellen Anwendung visualisiert.
Die Verfolgung konvektiver Zellen ermöglicht eine erweiterte Analyse der Zellen innerhalb von KONRAD. Unter anderem erfolgt mittels eines Warnindikators ein Hinweis zu Starkregen bei Überlappung der von einer verfolgten Primärzelle überdeckten Regionen zu mehreren aufeinanderfolgenden Detektionszeitpunkten. In diesen Fällen können sich aufgrund der langsamen Verlagerungsgeschwindigkeit der Zellen in kurzer Zeit lokal hohe Niederschlagsmengen akkumulieren. Der Warnindikator für Hagel steht im Zusammenhang mit der Anzahl von Pixeln einer Primärzelle, die $Z \geq 55\,\mathrm{dBZ}$ überschreiten (Mason, 1971).

Wie bei allen Zellverfolgungsalgorithmen treten auch in KONRAD bisweilen unvermeidbare Fehlzuordnungen auf, beispielsweise wenn das Verfahren im Vorprodukt eine falsche Zelle als Vorgängerin auswählt, die in der Nähe der korrekten Vorgängerzelle lag. Solche Fehlzuordnungen haben direkte Auswirkungen auf die den Primärzellen zugeschriebenen Lebenszyklen (s. Kapitel 4.3.2). Details zur Zellverfolgung in KONRAD finden sich in Lang et al. (2003). Für detaillierte Informationen zu weiteren radarbasierten Zellverfolgungsalgorithmen sei auf die Arbeiten zu den Verfahren TREC bzw. COTREC (Rinehardt und Garvey, 1978; Li et al., 1995), TITAN (Dixon und Wiener, 1993), TRACE3D (Handwerker, 2002; Schmidberger, 2018) und Rad-TRAM (Kober und Tafferner, 2009) verwiesen.

Das VX-Produkt von KONRAD

Das von KONRAD erstellte VX-Produkt bildet die Datengrundlage für die Analyse konvektiver Zellen in der vorliegenden Arbeit. Wie das RX-Produkt, auf dem das KONRAD-Verfahren basiert, liegt es in einer zeitlichen Auflösung von 5 min vor. Beim VX-Produkt handelt es sich um ASCII-Dateien, die in verschiedene Abschnitte gegliedert sind. Nach einer Kopfzeile zur Identifikation einer Datei folgt eine tabellarische Auflistung der im KONRAD-Verfahren bestimmten Ausgabegrößen für die Primärzellen (z. B. die Identifikationsnummer der Primärzelle, Position des Zentroids, Fläche der Primärzelle). Die Auflistung beinhaltet all diejenigen Zellen als numerische Objekte, die KONRAD in den vergangenen 30 min detektierte. Jede Zeile gehört zu einer Primärzelle. Eine Zelle, die in diesem Zeitraum kontinuierlich detektiert wurde, taucht demnach insgesamt siebenmal in der Tabelle auf. Die Sortierung in der Tabelle orientiert sich am Detektionszeitpunkt und

der Identifikationsnummer der Zellen. Im Anschluss folgt eine Auflistung aller Radarstationen, deren Daten in das jeweilige Radar-Komposit eingeflossen sind. Darüber hinaus sind solche Stationen gekennzeichnet, die temporär keine Daten lieferten (s. o.). Damit ist jederzeit nachvollziehbar, wie die exakte Datengrundlage des RX-Produkts aussah, welche in der Konsequenz Einfluss auf die Zelldetektion und -verfolgung durch KONRAD sowie die erkannten Primärzellen haben kann (s. Kapitel 4.3.2). Nach der Auflistung der Radarstationen folgt eine Ergänzungsliste, in der weitergehende Informationen abgespeichert sind. Beispielsweise finden sich dort ergänzend zum oben beschriebenen Hagelindikator Informationen zu Hagelwarnungen aus einem fünfzehnminütlich vorliegenden 3D-Radarprodukt, die auf dem Kriterium nach Waldvogel et al. (1979) basieren (Lang et al., 2003). Außerdem sind dort die Positionen der Sekundärzellen aufgelistet, die innerhalb von KONRAD auch in die Abschätzung der allgemeinen mittleren Verlagerungsrichtung eingehen. Ebenfalls sind detaillierte Informationen über die exakte geografische Position von größeren Starkregengebieten abgespeichert.
Der vorliegende VX-Datensatz für den Untersuchungszeitraum der Sommerhalbjahre 2011 – 2016 ist nicht lückenlos: Für 158 Zeiträume mit jeweils einer Dauer zwischen 5 min und 24 h fehlen Dateien oder sind fehlerhaft. Letzteres kann beispielsweise eine fehlerhafte zeitliche Zuordnung sein, die nachträglich manuell korrigiert wurde. Dass Dateien komplett fehlen, ist meist nur für einen Zeitraum von 5 – 60 min der Fall.
Als Nachfolge von KONRAD wird momentan die Neuentwicklung des Systems KONRAD3D präoperationell getestet (Werner, 2020). Dieses Verfahren berücksichtigt Radarmessungen aus unterschiedlichen Höhenschichten. Außerdem kann es auf neu entwickelte Techniken zur Qualitätssicherung von Radardaten und zur quantitativen Niederschlagsabschätzung zurückgreifen sowie die Methodik zur Hydrometeorklassifikation ausnutzen (vgl. Kapitel 4.1.1).

4.2 Daten aus dem Modell COSMO

4.2.1 Kurzbeschreibung von COSMO

Das COSMO-Modell ist ein nicht-hydrostatisches, atmosphärisches Vorhersagemodell für ein räumlich begrenztes Teilgebiet der Erde (*Limited Area Model*; Schättler et al., 2019). Seine ursprüngliche Version mit dem Namen ‚Lokal Modell' (LM), welches Deutschland und Teile seiner Nachbarstaaten abdeckte und Vorhersagen für die kommenden 48 h berechnete, wurde vom DWD Ende der 1990er Jahre entwickelt und einen Monat vor der Jahrtausendwende gemeinsam mit dem damals neuen ‚Global Modell' (GME) für die operationelle numerische Wettervorhersage bereitgestellt (NWV; Schulz und Schättler, 2014). Knapp sechs Jahre später folgte die operationelle Einführung der Erweiterung des LM zum ‚Lokal Modell Europa' (LME), das unter anderem eine Vergrößerung des Modellgebiets auf nahezu ganz Europa und eine Verlängerung des Vorhersagezeitraums auf 78 h vorwies. Im Jahr 2007 wurde das NWV-System um das ‚Lokal Modell Kürzestfrist' (LMK) ergänzt, das in höherer Auflösung auf Deutschland fokussierte. Die Weiterentwicklung und Verbesserung des LM erfolgt im Rahmen des internationalen *Consortium for Small-Scale Modelling* (COSMO)[2]. Diesem Zusammenschluss gehören neben dem DWD und dem Geoinformationsdienst der Bundeswehr die nationalen Wetterdienste der Schweiz, Italiens und Griechenlands sowie viele weitere meteorologische Institutionen aus meist europäischen Ländern an, die jeweils eigene operationelle Anwendungen des COSMO-Modells betreiben. 2007 wurde beschlossen, das LM in COSMO umzubenennen: Das LMK, das hauptsächlich Deutschland abdeckt, hieß fortan COSMO-DE. Das nahezu ganz Europa abdeckende LME erhielt den Namen COSMO-EU. Mittlerweile nutzt der DWD das Vorhersagemodell ICON (*Icosahedral Nonhydrostatic [Model]*) sowohl zur globalen als auch

[2] Umfangreiche Informationen finden sich auf der Website: `http://cosmo-model.org`.

seit 2016 zur europaweiten Wettervorhersage (ICON-Europa). Damit wurde COSMO-EU von ICON-Europa als Regionalmodell abgelöst. COSMO-D2, eine weiterentwickelte Version von COSMO-DE vervollständigt zusammen mit den entsprechenden Verfahren der Datenassimilation das aktuelle operationelle NWV-System des DWD. Während der Anfertigung der vorliegenden Arbeit fand der schrittweise Übergang von COSMO-D2 zu ICON-D2 statt.

COSMO zeichnet eine hohe Flexibilität bezüglich seines Einsatzbereichs aus. So dient es neben der Berechnung von hochaufgelösten regionalen Wettervorhersagen auch der Untersuchung verschiedenster wissenschaftlicher Anwendungen und Fragestellungen auf unterschiedlichsten räumlichen und zeitlichen Skalen. Darüber hinaus wird es auf einer weiteren Entwicklungslinie der *Climate Limited-area Modelling-Community* (CLM) als COSMO-CLM für regionale Klimasimulationen eingesetzt (Rockel et al., 2008).

Die horizontale Auflösung von COSMO in der operationellen DWD-Routine betrug im Untersuchungszeitraum rund 7 km (0,0625 °, EU) bzw. etwa 2,8 km (0,025°, DE). Damit erfasst COSMO auch nicht-hydrostatische Effekte auf Skalen der Größenordnungen $\mathcal{O}(10\,\text{km}) - \mathcal{O}(100\,\text{km})$. Während mit COSMO-EU im Vergleich zu den Globalmodellen bessere Vorhersagen unter anderem der bodennahen Wetterbedingungen, wie beispielsweise von Nebel, frontalen Niederschlägen und orografisch und thermisch induzierten lokalen Windsystemen im Fokus standen, zielte die hohe Auflösung von COSMO-DE besonders auf die direkte Simulation von hochreichender Feuchtkonvektion und deren Begleiterscheinungen ab (Schulz und Schättler, 2014; Baldauf et al., 2016). Die Modellgleichungen des COSMO-Modells beruhen auf den fundamentalen Bewegungsgleichungen für nicht-hydrostatische, kompressible Strömungen ohne Skalenapproximationen (vgl. Kapitel 2.1.2; z. B. Vallis, 2017; Schättler et al., 2019) und sind um weitere Tendenzgleichungen für den Wasserdampfgehalt (spezifische Feuchte) sowie den spezifischen Flüssigwassergehalt (Wolken- und Regentropfen) und den spezifischen Gehalt gefrorenen Wassers (Eis, Graupel, Schnee; jeweils als $q^{(i)}$ bezeichnet)

erweitert. Damit ergibt sich folgendes geschlossenes, nicht-lineares, gekoppeltes Gleichungssystem (Doms und Baldauf, 2018):

$$\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = -\nabla p + \rho \mathbf{g}' - 2\rho \boldsymbol{\Omega} \times \mathbf{v} - \rho \nabla \cdot \mathscr{T} \tag{4.5}$$

$$\frac{\mathrm{D}p}{\mathrm{D}t} = -\frac{c_p}{c_v} p \nabla \cdot \mathbf{v} + \left(\frac{c_p}{c_v} - 1 \right) Q_h + \frac{c_p}{c_v} Q_m \tag{4.6}$$

$$\rho c_p \frac{\mathrm{D}T}{\mathrm{D}t} = \frac{\mathrm{D}p}{\mathrm{D}t} + Q_h \tag{4.7}$$

$$\rho \frac{\mathrm{D}q^{(i)}}{\mathrm{D}t} = -\nabla \cdot \mathbf{J}^{(i)} + S^{(i)} \tag{4.8}$$

$$\rho = p \left[R_d (1 + \beta) T \right]^{-1} \ . \tag{4.9}$$

Darin steht Q_h für die diabatische Wärmeproduktion pro Einheitsvolumen, Q_m für den Einfluss von Konzentrationsänderungen der Wasserbestandteile $q^{(i)}$, $\mathbf{J}^{(i)}$ für den Diffusionsfluss der $q^{(i)}$, $S^{(i)}$ für deren Quellen und Senken und β für einen erweiterten Term für die Anteile der verschiedenen Wasserphasen in der Definition der virtuellen Temperatur. In COSMO sind der Gehalt gefrorenen Wassers sowie der Flüssigwassergehalt lediglich optionale prognostische Variablen, sodass das Modell auch mit vereinfachten wolkenmikrophysikalischen Prozessen anwendbar ist.

Die Gleichungen (4.5)–(4.9) werden aufgrund der näherungsweisen Kugelgeometrie der Erde in Kugelkoordinaten formuliert. Zur Vermeidung des sogenannten Polproblems erfolgt eine Reduzierung der Drängung der Meridiane mittels einer Rotation des Koordinatensystems. Für die Gitter der (ehemals) operationellen COSMO-Modelle wird die Lage der Pole dergestalt gekippt, dass der rotierte 0°-Meridian mit dem geografischen 10° O-Meridian übereinstimmt und der rotierte Äquator etwa von Schottland über Südschweden und das Baltikum nach Russland verläuft. Zudem erfolgt die Subtraktion eines horizontal homogenen (zeitlich invarianten, hydrostatisch balancierten) Grundzustands, sodass anstelle des

Drucks p die Fluktuationen dessen um einen höhenabhängigen Referenzdruck $p'(\mathbf{x},t) = p(\mathbf{x},t) - p_0(z)$ als prognostische Variablen dienen. Eine Transformation der Vertikalkoordinate in ein geländefolgendes Koordinatensystem ist implementiert, um numerische Probleme in orografisch gegliederten Regionen zu umgehen. Zur numerischen Näherungslösung der Modellgleichungen ist eine zeitliche und räumliche Diskretisierung erforderlich. Die horizontale Anordnung der Modellvariablen erfolgt gemäß des Arakawa-C-Gitters, während die vertikale Anordnung routinemäßig seit 2010 mittels einer modifizierten Version der Gal-Chen-Koordinate vorgenommen wird (Gal-Chen und Somerville, 1975; Arakawa und Lamb, 1977). Im operationellen Betrieb von COSMO-EU und COSMO-DE findet für die Zeitintegration das Verfahren von Klemp und Wilhelmson (1978b) Anwendung, welches unterschiedliche Zeitschritte für langsame und schnelle Moden einer kompressiblen Strömung erlaubt. Weitere Details finden sich bei Doms und Baldauf (2018).
Prozesse, die unterhalb der aufgelösten Skala ablaufen (sogenannte subskalige Prozesse), modelliert COSMO durch einen Satz von physikalischen Parametrisierungen (Doms et al., 2018). Dazu gehören irreversible, reibungsbedingte sowie diabatische Prozesse. Beispielsweise erfolgt eine Parametrisierung der Turbulenz in der freien Atmosphäre, wie auch eine für flache und hochreichende Konvektion. Deren Implementierung folgt dem Massenfluss-Schema nach Tiedtke (1989) oder dem erweiterten, zurzeit im Integrierten Vorhersagesystem des *European Centre for Medium-Range Weather Forecasts* (ECMWF) implementierten Tiedtke-Bechtold-Schema (Bechtold et al., 2001). Als weitere Parametrisierungen zu nennen sind die der Wolkenmikrophysik, der subskaligen Bewölkung, der subskaligen Orografie, von Süßwasserseen und Meereis, sowie des kurz- und langwelligen Strahlungstransfers (inklusive voller Rückkopplung mit den Wolkenschemata; Ritter und Geleyn, 1992). Zudem findet ein Boden- und Vegetationsmodell sowie eine Parametrisierung der Oberflächenflüsse Anwendung.

4.2.2 Datenassimilation für COSMO

Der operationelle Ablauf am DWD ist unterteilt in den Datenassimilationszyklus und die Erstellung der Hauptlaufanalysen und -vorhersagen. Die Hauptlaufanalysen von COSMO-EU beispielsweise wurden bis zum Ende der operationellen Verwendung am 1. Dezember 2016 nur für die Termine 00, 06, 12 und 18 UTC erstellt, an die sich die Berechnung der Vorhersage anschloss. Assimiliert wurden dort nur solche Daten, die bis 2 h 14 min nach dem Zeitpunkt der zu erstellenden Analyse vorlagen (Datenredaktionsschluss). Somit war beispielsweise eine 00 UTC-Hauptlaufanalyse gegen 2:30 UTC verfügbar. Im Datenassimilationszyklus wurden für jede volle Stunde Assimilationsanalysen in Blöcken von je drei Stunden mit einem späteren Datenredaktionsschluss erstellt. Die 00 – 02 UTC-Assimilationsanalysen waren daher erst gegen 5:10 UTC verfügbar, jedoch gingen mehr Beobachtungen in die Assimilation ein als bei der Erstellung der Hauptlaufanalysen. Dadurch erhoffte man sich eine Steigerung der Analysequalität.
Zur Bereitstellung eines skalenadäquaten Anfangszustands für die Vorhersage wurde als Analyseverfahren die sogenannte *Nudging*-Methode (auch: Newton'sche Relaxation) angewendet (Schraff, 1996, 1997; Schraff und Hess, 2013). Das *Nudging* stellt eine kontinuierliche vierdimensionale Datenassimilation dar, welche die prognostischen Modellvariablen während der Vorwärtsintegration des Modells innerhalb eines vorher festgelegten Zeitfensters an die Beobachtungen heranzieht (z. B. Davies und Turner, 1977; Stauffer und Seaman, 1990). Dies geschieht mit Hilfe eines additiven Zusatzterms in den jeweiligen Tendenzgleichungen, welche somit folgende Form für eine Prognosevariable ξ aufweisen:

$$\frac{\partial}{\partial t}\xi(\mathbf{x},t) = \mathscr{M}(\xi,\mathbf{x},t) + c_{n,\xi}\sum_{i=1}^{N_{obs}} w_i\left[\xi_i - \xi(\mathbf{x}_i,t)\right] \ . \tag{4.10}$$

Darin stellt $\mathscr{M}$ das reine Vorhersagemodell dar, also die Modelldynamik sowie die physikalischen Parametrisierungen. $c_{n,\xi}$ ist der sogenannte *Nudging*-Koeffizient. Die mit diesem multiplizierte Summe auf der rechten Seite von Gleichung (4.10) umfasst die Differenz aller Beobachtungswerte ξ_i in der Nähe des Modell-Gitterpunkts $\mathbf{x}$ zum Modellwert $\xi(\mathbf{x}_i, t)$ am jeweiligen Beobachtungsort $\mathbf{x}_i$ zum Zeitpunkt t. Obwohl das *Nudging* auf direkte Beobachtungen und nicht auf aus Beobachtungen abgeleitete gerasterte Beobachtungsanalysen angewendet wird und die Beobachtungen in der Regel nicht auf Modellgitterpunkten verortet sind, wird zur Bestimmung von $\xi(\mathbf{x}_i, t)$ aufgrund der generell mit COSMO verwendeten hohen Auflösung auf eine Interpolation der Werte verzichtet (Schraff und Hess, 2013). Die Prognosevariablen ξ, die das *Nudging* modifiziert, sind der Horizontalwind $\mathbf{u}$, die potentielle Temperatur θ und die relative Feuchte RH (auf allen Modellhauptflächen [Schichtmitten]) sowie der Luftdruck p (auf der untersten Modellhauptfläche; Schulz und Schättler, 2014). In der Praxis ist der Beitrag des *Nudgings* in der Regel kleiner als der jeweils größte Term in der Modelldynamik, d. h. das dynamische Gleichgewicht des Modells wird nicht übermäßig gestört. Unter Vernachlässigung von $\mathscr{M}$ in Gleichung (4.10) führt das *Nudging* mit dem üblicherweise verwendeten Wert $c_{n,\xi} = 6 \cdot 10^{-4}\,\mathrm{s}^{-1}$ zu einer exponentiellen Relaxation eines Modellwerts an einen Beobachtungswert innerhalb von etwa einer halben Stunde.

Im Untersuchungszeitraum fand darüber hinaus für das hochaufgelöste COSMO-DE (und seit dem 3. September 2014 auch für COSMO-EU) ein spezielles *Nudging*-Verfahren, das *Latent Heat Nudging*, Anwendung, welches hochaufgelöste Daten der aus dem Reflektivitätsfaktor Z abgeleiteten Regenrate R in das Modell assimiliert[3] (vgl. Kapitel 4.1.1;

[3] Für Bereiche außerhalb des Radarverbunds des DWD wurde hierfür während des Untersuchungszeitraums auf den weite Teile Europas abdeckenden OPERA-Datensatz zurückgegriffen: `https://www.eumetnet.eu/activities/observations-programme/current-activities/opera/`.

Stephan et al., 2008; Baldauf et al., 2016; DWD, 2021a). Über das Einbringen von zusätzlichen Temperatur- und Feuchteinkrementen beeinflusst dieses *Nudging* die Modelldynamik dergestalt, dass sich der Modellniederschlag an die Beobachtung angleicht. Seit März 2017 findet für die Datenassimilation in COSMO-DE bzw. in seinen Nachfolger COSMO-D2 das neu entwickelte System KENDA (*Kilometer-Scale Ensemble Data Assimilation*; Schraff et al., 2016) mit einem eingebauten LETKF (*Local Ensemble Transform Kalman Filter*; Hunt et al., 2007) Anwendung, in dem das *Latent Heat Nudging* zunächst weiterhin der Assimilation der Regenrate dient. Zurzeit sind Methoden in der Entwicklung, den Radarreflektivitätsfaktor Z sowie die Radialgeschwindigkeit mit Hilfe des Radar-Vorwärts-Operators EMVORADO (*Efficient Modular Volume Scanning Radar Operator*) durch KENDA direkt zu assimilieren (Zeng, 2013; Zeng et al., 2016; Blahak und de Lozar, 2020; DWD, 2021c).

4.2.3 Assimilationsanalysen von COSMO-EU

Assimilationsanalysen von COSMO-EU stehen in der Datenbank des DWD in stündlicher Auflösung zur Verfügung (vgl. Kapitel 4.2.1). Da ab 2016 das neue ICON-Europa das alte COSMO-EU im operationellen Routinebetrieb abgelöst hat, liegen Analysen von COSMO-EU bis einschließlich November 2016 vor (GRIB2-Format). Der Datensatz ist nahezu vollständig. Lediglich vier der 26 352 verwendeten Analysedateien im Untersuchungszeitraum der Sommerhalbjahre 2011 – 2016 sind fehlerhaft.
Das Modellgebiet von COSMO-EU erstreckt sich über nahezu ganz Europa (Abbildung 4.3) mit 665 $\times$ 657 strukturiert verteilten horizontalen Gitterpunkten mit einer Gitterpunktsdistanz von 0,0625 ° (vgl. Kapitel 4.2.1). Vertikal sind 40 Modellhauptflächen definiert, die sich von 10 m über Grund bis in eine Höhe von 21,75 km erstrecken. Je weiter zwei Hauptflächen von

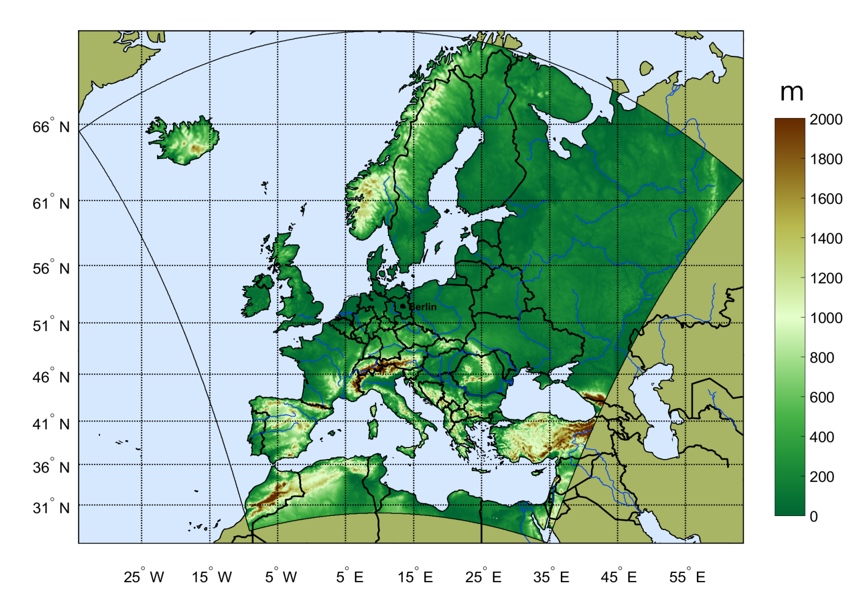


Abbildung 4.3: Erstreckung des Modellgebiets von COSMO-EU inklusive der Modellorografie (m ü. NN).

der Erdoberfläche entfernt sind, desto größer ist der Abstand zwischen ihnen (Schulz und Schättler, 2014; Doms und Baldauf, 2018). Die Assimilationsanalysen von COSMO-EU liegen sowohl auf diesem Modellgitter als auch auf interpolierten Höhen- und Druckschichten vor.

4.3 Methoden der Datenaufbereitung

Das Ziel der Datenaufbereitung ist die Erstellung eines geeigneten Datensatzes für den Untersuchungszeitraum der Sommerhalbjahre 2011 – 2016, der Lebenszyklen konvektiver Zellen sowie Informationen über die zugehörigen atmosphärischen Umgebungsbedingungen beinhaltet. Auf diesem Datensatz basieren alle weiterführenden Analysen der Zellattribute und Umgebungsvariablen (Kapitel 5) sowie die Entwicklung von Vorhersageverfahren,

die verschiedene Zellattribute anhand der vorherrschenden Umgebungsbedingungen und der Zellhistorie als gewinnbringende Information für *Nowcasting*-Verfahren abschätzen können (Kapitel 6). Erforderlich ist dafür, dass die Lebenszyklen der Primärzellen von KONRAD reale konvektive Entwicklungen abbilden. Wie später dargelegt, ist dies für viele der abgespeicherten Primärzellen jedoch nicht der Fall. Daher ist nach der Zusammenstellung aller Lebenszyklen auf der Basis des VX-Produkts von KONRAD (Kapitel 4.3.1) eine umfangreiche Filterung der Daten nötig, um fehlerhafte Lebenszyklen auszusortieren (Kapitel 4.3.2). Die Umgebungsbedingungen während des Auftretens der konvektiven Zellen quantifizieren verschiedene Umgebungsvariablen, die auf COSMO-EU-Assimilationsanalysen basieren (Kapitel 4.3.3). Die Zusammenführung der gefilterten objektbezogenen Lebenszyklen und der auf einem Gitter flächendeckend vorliegenden Umgebungsvariablen zu einem kombinierten objektbezogenen Datensatz komplettiert die Datenaufbereitung (Kapitel 4.3.4).

4.3.1 Erstellung zusammenhängender Lebenszyklen aus den Daten des Zellverfolgungsalgorithmus KONRAD

Wie in Kapitel 4.1.2 beschrieben liegen die Daten der von KONRAD detektierten konvektiven Zellen in separaten Dateien für jeden Zeitpunkt der fünfminütlichen Radarmessungen vor. Somit ist eine Primärzelle bei längerer Lebensdauer in mehreren Dateien enthalten. Um die zeitliche Entwicklung der Attribute einer bestimmten Primärzelle leichter analysieren zu können, sollen die Informationen aus den unterschiedlichen Dateien für diese zusammengeführt werden, sodass ein numerisches Datenobjekt entsteht, welches den vollständigen Verlauf als zusammenhängenden Lebenszyklus enthält.
Die Entwicklung eines Verfahrens zur Erstellung zusammenhängender Lebenszyklen erfolgte für einen einmonatigen Testzeitraum, um den Rechenaufwand gering zu halten. Die Periode 27. Mai – 26. Juni 2016, in der ungewöhnlich viele konvektive Zellen über ganz Deutschland auftraten, die

Tabelle 4.1: Übersicht über die in der vorliegenden Arbeit verwendeten Zellattribute aus KONRAD inklusive Formelzeichen und Einheit (ZO = Zellobjekt).

Zellattribut	Zeichen	Einheit
geografische Länge eines Zentroids	λ_Z	°
geografische Breite eines Zentroids	ϕ_Z	°
Alter eines ZO	t	min
KONRAD-interne Identifikationsnummer des ZO	ID_Z	—
Zellfläche ($\hat{=}$ Anzahl der Zellpixel mit $Z \geq 46$ dBZ)	A_Z	km^2
Zellkernfläche ($\hat{=}$ Zellpixelanzahl mit $Z \geq 55$ dBZ)	$A_{Z,K}$	km^2
Azimut der Verlagerungsrichtung des ZO	α_Z	°
Verlagerungsgeschwindigkeit des ZO	c_Z	$\mathrm{m\,s^{-1}}$
geografische Länge des westlichen Rands des ZO	$\lambda_{Z,W}$	°
geografische Länge des östlichen Rands des ZO	$\lambda_{Z,O}$	°
geografische Breite des südlichen Rands des ZO	$\phi_{Z,S}$	°
geografische Breite des nördlichen Rands des ZO	$\phi_{Z,N}$	°

teils heftige Regenfälle mit schweren Überschwemmungen, Hagel und sogar Tornados verursachten (Piper et al., 2016), bietet dafür gute Voraussetzungen, da hier eine große Anzahl von Zellen unterschiedlicher Organisationsformen vorkam. Anschließend wird dieses Verfahren auf den gesamten Untersuchungszeitraum angewendet. Eine Anwendung ist bei entsprechender Berücksichtigung fehlender und fehlerhafter Dateien auf jeden beliebigen Zeitraum möglich, für den VX-Dateien vorhanden sind.

Im ersten Schritt werden die Daten aus den VX-Dateien so weit reduziert, dass aus jeder Datei nur die Informationen über die Attribute aller zum aktuellen Zeitpunkt registrierten Primärzellen sowie die Informationen über die Radarverfügbarkeit ausgeschnitten und in ASCII-Tabellen zwischengespeichert werden. Dabei können die verschiedenen gewünschten Zellattribute aus

Tabelle 4.2: Übersicht über vorhandene Radar- und Zeitinformationen, die anhand der VX-Dateien von KONRAD bestimmt werden können.

Beschreibung	Abkürzung
Indikator, welche Radardaten 5 min früher nicht verfügbar sind	RTM
Indikator, welche Radardaten zum jeweiligen Zeitpunkt nicht verfügbar sind	RTN
Indikator, welche Radardaten 5 min später nicht verfügbar sind	RTP
Datums- und Uhrzeitangabe	DAT

Tabelle 4.1 flexibel übertragen und mit den in Tabelle 4.2 aufgeführten Radar- und Zeitinformationen ergänzt werden. Im Folgenden dient der Begriff Zellobjekt als Bezeichnung für von KONRAD detektierte Primärzellen und ermöglicht die sprachliche Abgrenzung gegenüber beobachteten, realen konvektiven Zellen.

Für jeden Tag im Untersuchungszeitraum wird damit ein Datensatz erstellt, der alle Informationen über alle an diesem Tag registrierten Zellobjekte enthält. Eine Dimension steht für die Identifizierung des Zellobjekts, eine zweite Dimension spannt sich über alle Zeitpunkte der Detektionen auf, eine dritte Dimension gibt Raum für alle in den Tabellen 4.1 und 4.2 aufgelisteten Zellattribute und Informationen. Damit lässt sich einfach und schnell auf alle Attribute eines bestimmten Zellobjekts zu einem beliebigen Zeitpunkt des Lebenszyklus durch Indizierung zugreifen. Zellen, die sowohl vor als auch nach Mitternacht von KONRAD detektiert wurden, liegen mit ihrem vollen Lebenszyklus als Zellobjekte in der Datei des ersten Tags.

4.3.2 Filterung der Daten des Zellverfolgungsalgorithmus KONRAD

Trotz der ausgereiften Methodik zur Detektion und Verfolgung konvektiver Zellen in KONRAD treten für das Zellverfolgungsverfahren einige Schwierigkeiten auf. Diese resultieren sowohl aus der Dynamik der konvektiven Zellen als auch aus mathematisch-technischen Herausforderungen. Insbesondere ergeben sich neben fehlerhaften Dateien weitere zu berücksichtigende Problempunkte bei der Qualitätskontrolle der erstellten Lebenszyklen:

(a) Der Start- bzw. Endpunkt der Zelle kann außerhalb des durch den Radarverbund abgedeckten Gebiets liegen.

(b) Zellen können Gebiete überqueren, für die zum entsprechenden Zeitpunkt keine Radarprodukte vorhanden sind.

(c) Die Beschreibung eines Lebenszyklus im eigentlichen Sinn ist für Zellobjekte mit einer sehr kurzen Lebensdauer von weniger als etwa 20 min problematisch.

(d) Die Nomenklaturregeln von KONRAD, d. h. die Vorschriften für die Zuweisung der Identifikationsnummer ID_Z zu den Zellobjekten, müssen beachtet werden.

(e) Zuordnungsprobleme entstehen durch das Verschmelzen und Teilen von Zellen (*Merging* bzw. *Splitting*). Dabei ist physikalisch bereits nicht klar, wie der Lebenszyklus solcher Zellen zu definieren ist. Zudem kann die Handhabung der Zellobjekte durch den Algorithmus von KONRAD in solchen Fällen zu Problemen führen.

In der Atmosphäre kommt es je nach der Organisationsform der Konvektion häufig zu den in (e) erwähnten Verschmelzungen und Teilungen von Zellen (vgl. Kapitel 2.2). Manchmal entstehen aber auch ganz in der Nähe von bereits existierenden Zellen weitere Zellen, manchmal dissipiert eine Zelle in

der Nähe von anderen (vgl. Abbildung 4.2 zur Veranschaulichung von KONRAD-Primärzellen in einer Gewitterlinie). Diese Dynamik mit einem Zellverfolgungsalgorithmus zu erfassen, stellt Entwickler*innen generell vor viele Herausforderungen. Um eine möglichst realitätsnahe Stichprobe von konvektiven Zellen zu erhalten, findet eine Beschränkung auf isolierte Zellen, also auf Einzel- und Superzellen statt. Dadurch kann eine adäquate Filterung des Datensatzes den Problempunkt (e) größtenteils umgehen.
Für einige Testfälle, in denen Radarbilder auf Teilungen und/oder Verschmelzungen von Zellen hindeuten, wird die in (d) erwähnte entsprechende Zuweisung der Identifikationsnummern durch KONRAD genauer untersucht. In allen Fällen, in denen Multizellen oder ein MCS auftreten (vgl. Kapitel 2.2.2 und 2.2.4), kommen durch die Zuweisungen unrealistische Verläufe von Zellattributen zustande, die zur gleichen ID_Z gehören. Wapler (2021) beschreibt, dass im Fall einer Zellteilung entweder beide neu gebildeten Zellen eine neue ID_Z erhalten, KONRAD also beide als neu gebildete Zellobjekte behandelt, oder eine der beiden neuen Zellen die ID_Z des ursprünglichen Zellobjekts behält, während die andere eine neue ID_Z zugewiesen bekommt. Beides führt dazu, dass keines der Zellobjekte korrekt einen kompletten Lebenszyklus (mit einer entsprechenden Lebensdauer) widerspiegelt. Im ersten Fall können die beiden neuen Zellobjekte zum Zeitpunkt der (scheinbaren) ersten Detektion ($t = 2\,\text{min}$; vgl. Kapitel 4.1.2), die jedoch möglicherweise nicht mehr dem Cumulusstadium der Zellen zuzuschreiben ist, bereits eine große Zellfläche A_Z aufweisen (vgl. Kapitel 2.2.1). Im zweiten Fall kann für das Zellobjekt, das die ursprüngliche ID_Z behält, eine starke Abnahme der Zellfläche A_Z auftreten, sodass es so aussieht, als sei die entsprechende Zelle im Dissipationsstadium. Ähnliche Kausalketten lassen sich auch bei einer Verschmelzung zweier Zellen aufstellen. Daher ist eine Bereinigung des Datensatzes erforderlich.
Um all diese Aspekte zu berücksichtigen, erfolgt diese Bereinigung mit Hilfe von mehreren, neu entwickelten objektiven Filtermethoden in einer tageweisen und zugleich objektweisen Vorgehensweise. Für jedes Zellobjekt läuft der

Algorithmus konsekutiv durch alle Detektionszeitpunkte, wobei die Verlagerungsrichtung α_Z und -geschwindigkeit c_Z der Zellen dabei explizit aus den Veränderungen von λ_Z und ϕ_Z berechnet wird. Die Entwicklung der verschiedenen Filter und entsprechende Sensitivitätsuntersuchungen erfolgten auf der Basis des Testzeitraums 27. Mai – 26. Juni 2016. Im Anschluss wurde der gesamte Datensatz des Untersuchungszeitraums 2011 – 2016 der entwickelten Filterung unterzogen.

Prä- und Postfilterung

Da die Prozessierung allgemein für jede beliebige Zeitspanne mit frei wählbaren Start- und Endzeitpunkten anwendbar sein soll, dienen die Prä- und Postfilterung lediglich dazu, mit Hilfe der Datums- und Uhrzeitangabe DAT (Tabelle 4.2) diejenigen Zellobjekte auszusortieren, die zum Start- bzw. Endzeitpunkt der gewählten Zeitspanne von KONRAD registriert wurden.

Filterung hinsichtlich der Radarabdeckung

KONRAD kann den vollständigen Lebenszyklus einer konvektiven Zelle nur abbilden, wenn die Zelle zu jedem Zeitpunkt von einem Radar erfasst wird. Probleme entstehen, wenn innerhalb des Untersuchungsgebiets Radardaten fehlen (Problem (b)). Daher werden zunächst aussortiert: 1) Alle Zellobjekte, die KONRAD in einem Gebiet zum ersten Mal registriert, von dem 5 min zuvor keine Daten vorliegen. 2) Alle Zellobjekte, die KONRAD in einem Gebiet zum letzten Mal registriert, von dem 5 min danach keine Daten mehr vorliegen. 3) Alle Zellobjekte, die in ein (dauerhaft) datenloses Gebiet ziehen bzw. aus einem solchen hinausziehen. Zur numerischen Umsetzung sind die in Tabelle 4.2 aufgelisteten Radarinformationen RTM, RTN und RTP neben der Position der Zentroide (λ_Z, ϕ_Z) unabdingbar.

Bei der Durchführung dieser Filterung helfen Gebietsmasken (kurz: Masken), die Regionen überspannen, welche im Fall fehlender Daten bestimmter Radare von keinem Radar abgedeckt werden. Solche Masken sind für acht verschiedene Konstellationen von operationell arbeitenden Radaren notwendig, die im Untersuchungszeitraum 2011 – 2016 für bestimmte Zeiträume den Radarverbund bildeten[4] (vgl. Kapitel 4.1.1). Die Masken haben eine Auflösung von $\Delta\lambda = \Delta\phi = 0{,}04\,°$, meridional also von etwa $\Delta y \approx 4{,}5$ km und zonal breitengradabhängig zwischen $\Delta x \approx 2{,}5$ und 3,1 km. Um eine schmale Pufferzone um die vom DWD angegebene Reichweite der Einzelradarmessungen zu generieren, wird hier als Reichweite ein Radius von $R_{Radar} = 145$ km angenommen (vgl. Kapitel 4.1.1). In Abbildung 4.4 sind beispielhaft zum einen die Maske für einen Datenausfall des Radars Neuheilenbach in der Eifel (orangefarbene und rote Kreuze) und zum anderen die Maske der Randpunkte des Radarverbunds (grüne Kreuze) dargestellt. Die Unterscheidung in Punkte der Maske eines Gebiets (orange) und Randpunkte dieser Maske (rot) beschleunigt den Algorithmus. Um bei der oben erwähnten Auflösung genügend Randpunkte zu haben, ist eine Breite des Rands von 9 km $\approx 2\Delta y$ sinnvoll. Im Fall, dass vom Radar am Standort Neuheilenbach keine Daten verfügbar sind, ist das rot-orange eingefärbte Gebiet (Teile von Luxemburg, Lothringen, des Saarlands und weitere) nicht abgedeckt.
Die Filterkriterien sind derart gestaltet, dass ein Zellobjekt genau dann aussortiert wird, wenn sich sein Zentroid zu nahe an einem Punkt der für den jeweiligen Zeitpunkt (bzw. 5 min früher oder später) gültigen Maske befindet, deren Auswahl jeweils über *DAT* und *RTN* (bzw. *RTM* oder *RTP*) erfolgt (im Folgenden als Teilfilter 1 – 3 bezeichnet). Ausgehend von einer maximalen realistischen Verlagerungsgeschwindigkeit einer Zelle von

[4] Die Zeiträume, in denen Ausfallsicherungsradare bestimmte Radare ersetzten, behandelt die Filterung aus Gründen der Komplexität so, als ob die regulären Radare an ihren jeweiligen Standorten, die nur unweit der Standorte der Ausfallsicherungsradare liegen, verwendet worden wären.

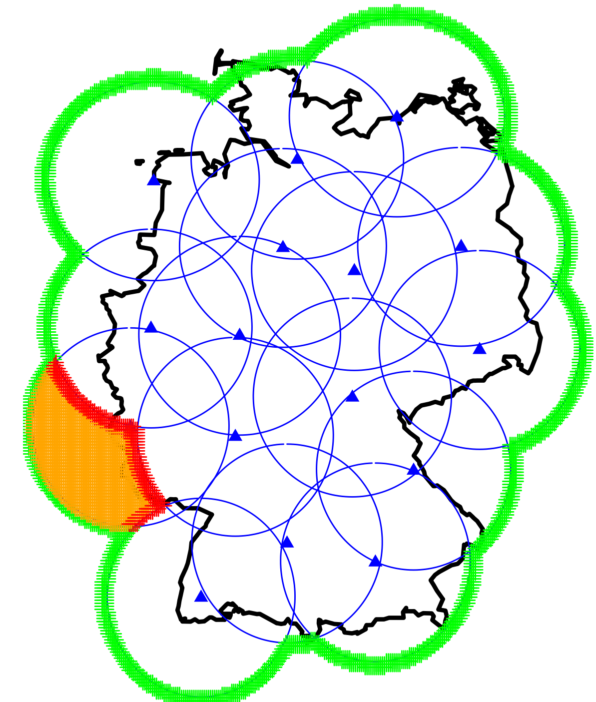


Abbildung 4.4: Zur Illustration der in mehreren Filtern verwendeten Gebietsmasken. Die Standorte der 2011 operationellen Radare des DWD-Radarverbunds sind mit blauen Dreiecken, deren radiale Reichweiten (hier: 145 km) mit blauen Kreisen gekennzeichnet. Siehe Fließtext für eine ausführliche Erläuterung der farbigen Markierungen.

$c_{Z,max} = 150\,\mathrm{km\,h^{-1}}$ gilt als guter Wert für den minimal zugelassenen Abstand eines Zellobjekts zu einem Punkt bzw. Randpunkt einer Maske $d_{krit,Maske} = c_{Z,max} 12^{-1}\,\mathrm{h} + \max(\Delta x, \Delta y) \approx 17\,\mathrm{km}$. Dabei wird verwendet, dass die VX-Daten in fünfminütlicher Auflösung vorliegen. Der Wert von $c_{Z,max}$

wird aus dem Grund etwas höher als im Zellverfolgungsverfahren von KONRAD angesetzt ($c_{Z,max} = 110\,\mathrm{km\,h^{-1}}$ plus mögliche Verlagerung des Zentroids innerhalb einer Primärzelle; vgl. Kapitel 4.1.2), da in mesoskaligen Fronten eingelagerte konvektive Zellen durchaus mit sehr hohen Geschwindigkeiten ziehen können, wie beispielsweise während des Orkans Kyrill am 18. Januar 2007 (Fink et al., 2009). Die Umsetzung dieser Filterung erfolgt, indem der Algorithmus in einem quasi-rechteckigen Umfeld mit einer Kantenlänge von je 0,5° um das Zentroid eines Zellobjekts nach den Punkten bzw. Randpunkten der momentan relevanten Maske sucht. Im Fall eines Treffers berechnet er dann im zweiten Schritt erst die einzelnen Distanzen in Kugeloberflächengeometrie und vergleicht diese anschließend mit dem minimal zugelassenen Abstand $d_{krit,Maske}$. Ist eine dieser Distanzen kleiner als $d_{krit,Maske}$, so befindet sich das betrachtete Zellobjekt zu nahe an einem Punkt oder Randpunkt der Maske und wird daher aussortiert.
Solange lediglich von einem oder zwei Radaren zu einem Zeitpunkt keine Daten vorliegen, sind die beschriebenen Filter wirksam. Sobald dies für drei oder mehr Radare der Fall ist, sortiert die Filterung alle zu diesem Zeitpunkt vorhandenen Zellobjekte aus, da der Rechenaufwand für die Kombination der Gebietsmasken stark mit der Anzahl von Radaren mit fehlenden Daten steigt (Teilfilter 4). Während der Sommerhalbjahre 2011 – 2016 waren zu etwa 11,6 % aller Zeitpunkte Daten von drei oder mehr Radaren gleichzeitig nicht verfügbar. Die Filterung sortiert in diesem Zuge auch Zellobjekte aus, die KONRAD zu einem Zeitpunkt 5 min früher oder später gegenüber einem Zeitpunkt registrierte, zu dem keine oder eine fehlerhafte VX-Datei vorliegt.

Filterung am Rand des Verbundgebiets

Wie eingangs dieses Unterkapitels in der Problemstellung (a) dargelegt ist die Berücksichtigung aller Zellen, die in das Gebiet des Radarverbunds hinein- oder aus ihm hinausziehen, zur Analyse von vollständigen Lebenszyklen ebenfalls nicht geeignet. Dazu setzt auch dieser Filter unter Verwendung von

λ_Z und ϕ_Z ein rechteckiges Umfeld um die Zentroide der Zellobjekte ein, um nach Punkten in der Maske der Randpunkte des Radarverbunds zu suchen, die innerhalb dieses Umfelds liegen. Im Fall eines oder mehrerer Treffer berechnet der Filter im zweiten Schritt wieder die jeweiligen Distanzen der Position des Zentroids von den Randpunkten und vergleicht diese mit dem minimal zugelassenen Abstand $d_{krit,Maske}$. Ist eine dieser Distanzen kleiner als $d_{krit,Maske}$, so befindet sich das betrachtete Zellobjekt zu nahe an einem der Randpunkte des Radarverbunds und wird daher aussortiert.

Filterung von falschen Zuweisungen

Das in Kapitel 4.3.1 beschriebene Verfahren setzte die Lebenszyklen der von KONRAD registrierten Zellobjekte für jeden Tag anhand ihrer Identifikationsnummer ID_Z zusammen, da jede ID_Z während einer konvektiven Wetterlage nur einmal vergeben wird. Es stellte sich jedoch heraus, dass zwei gänzlich unabhängige Zellobjekte in den Daten an einem Tag vereinzelt dieselbe Identifikationsnummer ID_Z tragen. Sehr selten hat sogar ein Zellobjekt zu einem Zeitpunkt eine bestimmte ID_Z und 5 min später ein anderes Zellobjekt, das teils mehrere 100 km entfernt ist, dieselbe ID_Z. In diesen Fällen setzen sich folglich die erstellten Lebenszyklen aus zwei verschiedenen Zellobjekten zusammen. Anhand des genauen Detektionszeitpunkts DAT und der Information über die Position der Zentroide (λ_Z, ϕ_Z) lassen sich solche Fehlzuweisungen herausfiltern.

Filterung kurzer Lebenszyklen

Zellobjekte, die KONRAD nur einmal registrierte, werden unter Verwendung des Alters t aussortiert. Dieses Vorgehen lässt sich, wie in der Problemstellung (c) angedeutet wurde, prinzipiell auf Zellen erweitern, die nur wenige Male detektiert wurden. Dabei stellt sich die Frage, bei welcher Lebensdauer genau die Grenze zu ziehen ist. Da KONRAD Zellen nur detektiert, wenn auf einer zusammenhängenden Fläche von mindestens 15 km^2 ein Schwellenwert

von $Z = 46$ dBZ erreicht bzw. überschritten wird, können Teile des Cumulus- und des Dissipationsstadiums einer konvektiven Zelle nicht abgebildet werden (vgl. Kapitel 4.1.2). Einer konvektiven Zelle lässt sich daher von der ersten Wolkenbildung bis zur Dissipation bereits eine reale Lebensdauer von mindestens etwa $30 - 45$ min attestieren, auch wenn sie nur zweimal in Folge das KONRAD-Kriterium überschreitet. Dies lässt sich insbesondere für nicht allzu starke Einzelzellen häufig beobachten (vgl. Kapitel 2.2.1). Aus diesem Grund behält der Filter alle Zellobjekte mit zwei oder mehr Detektionen bei.

Clusterfilter

Der Clusterfilter soll den Datensatz von Zellobjekten bereinigen, die potentiell aufgrund von Verschmelzungen oder Teilungen von Zellen keinen adäquaten Lebenszyklus vorweisen (s. o.). Er setzt die Forderung um, dass alle Zellobjekte, die zum ersten bzw. letzten Zeitpunkt ihres Lebenszyklus zu nahe an einem anderen Zellobjekt (einer Nachbarzelle, kurz: einem Nachbarn) sind, aussortiert werden müssen (Teilkriterium A). Auch der Nachbar selbst muss aussortiert werden, da er potentiell aus einer Zellteilung mit dem anderen Objekt heraus entstanden ist oder mit dem Zellobjekt im nächsten Zeitschritt als ein gemeinsames Objekt erscheint (Teilkriterium B). Jedes Zellobjekt kann gleichzeitig mehrere Nachbarn haben und selber für mehrere Objekte ein Nachbar sein. Nicht jeder identifizierte Nachbar ist jedoch automatisch mit einer realen Verschmelzung oder Aufteilung in Verbindung zu bringen. Der Clusterfilter bietet daher keine optimale Lösung für die in (e) formulierte Problemstellung. Es ist jedoch eine valide Annahme, dass dieser Filter den Datensatz von Zellobjekten bereinigt, deren zugehörige Attribute unrealistische Verläufe aufweisen, und ihn damit auf eine Teilrepräsentation des konvektiven Spektrums beschränkt.

Auch dieser Filter sucht in einem rechteckigen Umfeld der Kantenlänge 0,5° um das Zentroid unter Verwendung von λ_Z und ϕ_Z sowie von *DAT* nach potentiellen Nachbarn. Finden sich weitere Zellobjekte im Umfeld, berechnet

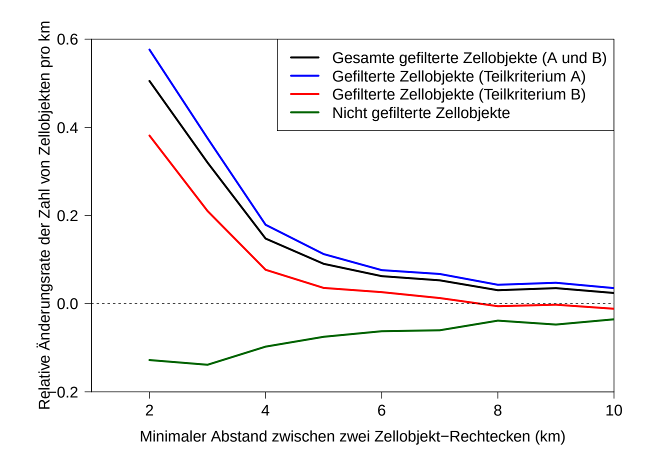


Abbildung 4.5: Sensitivitätsuntersuchung zum Clusterfilter im Testzeitraum 27. Mai – 26. Juni 2016. Auf der Abszisse ist der minimal erlaubte Abstand zwischen zwei Zellobjekten $d_{krit,Nachbar}$ aufgetragen. Die Ordinate gibt an, wie stark sich die Anzahl bestimmter Zellobjekte aufsteigend vom einen zum nächsten Abszissenwert (Intervall: 1 km) relativ ändert.

der Filter unter Verwendung des von $\lambda_{Z,W}$, $\lambda_{Z,O}$, $\phi_{Z,S}$ und $\phi_{Z,N}$ aufgespannten, die Zellfläche umrahmenden Rechtecks die jeweiligen Distanzen zwischen den nächstgelegenen Rändern der Zellobjekte (vgl. Kapitel 4.1.2). Ist eine dieser Distanzen kleiner als der minimal erlaubte Abstand $d_{krit,Nachbar}$ (s. u.), so befindet sich das betrachtete Zellobjekt zu nahe an einem der anderen Objekte, weshalb beide aussortiert werden.

Eine Sensitivitätsuntersuchung dient der Bestimmung eines geeigneten Schwellenwerts für den minimal erlaubten Abstand $d_{krit,Nachbar}$, den die Ränder zweier Zellobjekte zu einem relevanten Zeitpunkt haben dürfen (Abbildung 4.5). Dazu durchlaufen die Zellobjekte aus dem Testzeitraum 27. Mai – 26. Juni 2016 die Prä- und Postfilterung, die Filterung hinsichtlich der Radarabdeckung, die Filterung am Rand des Verbundgebiets, die Filterung von falschen Zuweisungen sowie diejenige kurzer Zelldetektionen und den Clusterfilter konsekutiv durch alle Detektionszeitpunkte. Die Anzahl der

nicht aussortierten Objekte (grüne Linie in Abbildung 4.5), sinkt wie erwartet mit steigenden Werten von $d_{krit,Nachbar}$. Die Abnahme geschieht allerdings immer langsamer. Umgekehrt nimmt die Anzahl der durch den Clusterfilter gefilterten Objekte (schwarze Linie) mit steigendem $d_{krit,Nachbar}$ zu. Ab $d_{krit,Nachbar} \approx 4-5\,\mathrm{km}$ ist die Zunahme deutlich geringer. Daher ist der Kompromiss $d_{krit,Nachbar} = 5\,\mathrm{km}$ als Filterparameter eine vertretbare Wahl, um möglichst viele Fälle von potentiellen Verschmelzungen und Teilungen konvektiver Zellen zu eliminieren und gleichzeitig nicht zu viele Fälle der Neubildung oder Dissipation einer Zelle in der näheren Umgebung von bereits existierenden Zellen zu unterdrücken. Die positive Änderungsrate aller durch den Clusterfilter aussortierten Zellobjekte für größere Werte von $d_{krit,Nachbar}$ und die langsame asymptotische Annäherung der Kurve an die Null ist ein Indiz dafür, dass solche Neubildungen oder Dissipationen in der Umgebung anderer Zellen durchaus auftreten.

Filter für die Zellfläche und die Verlagerungsrichtung

Als Ergänzung zum Clusterfilter dienen Filter für die Zellfläche und die Verlagerungsrichtung dazu, unrealistische Entwicklungen dieser Zellattribute zu identifizieren, welche auf falsche Zuweisungen der Zellobjekte hindeuten. Diese fußen auf dem allgemeinen Wissen über den Lebenszyklus konvektiver Zellen (Kapitel 2.2) sowie einer genaueren Analyse der Entwicklungen einiger prominenter Fallbeispiele von starken konvektiven Ereignissen (nicht gezeigt).

Der erste dieser Filter sortiert alle Zellobjekte aus, für die zum Zeitpunkt der ersten oder letzten Detektion A_Z Werte von mehr als $A_{Z,krit} = 40\,\mathrm{km}^2$ besitzt. Dadurch werden Zellen aussortiert, die potentiell vor der ersten Detektion bzw. nach der letzten Detektion durch KONRAD in der Realität (bereits) eine große Ausdehnung hatten, durch KONRAD aufgrund einer unbekannten Ursache jedoch nicht erkannt wurden. Des Weiteren ist binnen eines Zeitschritts

von 5 min eine maximale Änderung der Zellfläche von $50\,\text{km}^2$, d. h.

$$\left|\left(\frac{\mathrm{d}A_Z}{\mathrm{d}t}\right)\right|_{krit} = 10\,\text{km}^2\,\text{min}^{-1} \tag{4.11}$$

erlaubt. So sollen als unrealistisch einzustufende Sprünge in der Entwicklung der Zellfläche innerhalb des Lebenszyklus der Zellobjekte erkannt werden.
Der zweite Filter sortiert alle Zellobjekte aus, deren Verlagerung des Zentroids α_Z für mindestens einen Zeitschritt von 5 min eine Richtungsänderung aufweist, die größer als ein Schwellenwert $\alpha_{Z,krit}$ ist. Der Filter glättet zuvor α_Z über drei Zeitpunkte, um den Einfluss der internen Verlagerung des Zentroids bezüglich des Zellobjekts zu reduzieren (vgl. Kapitel 4.1.2). Der Filter findet nur Anwendung, wenn sich das Zentroid innerhalb der letzten 5 min um eine bestimmte Strecke verlagert hat. Der Grund dafür ist, dass (quasi-)stationäre Zellen große Änderungen der schwer festzulegenden Verlagerungsrichtung aufweisen können, dies allerdings nur auf kurzer Distanz. Eine Sensitivitätsuntersuchung für den Testzeitraum zeigt, dass analog zur Clusterfilterung die Wahl der Schwellenwerte nur einen Kompromiss zwischen der Filterung realistischer und unrealistischer Entwicklungen darstellen kann, da es nicht möglich ist, alle einzelnen Zellobjekte im Detail zu überprüfen (Abbildung 4.6). In dieser Untersuchung werden dieselben Filter wie in Abbildung 4.5 verwendet (der Clusterfilter entsprechend mit $d_{krit,Nachbar} = 5\,\text{km}$) und der Filter für die Verlagerungsrichtung mit verschiedenen Kombinationen der Schwellenwerte hinzugefügt. In die Festlegung von $(c_{Z,krit}, \alpha_{Z,krit})$ auf die Werte $(12\,\text{km}\,\text{h}^{-1}, 30°)$ als am besten geeignete Kombination fließt das Wissen über die Auflösung der Radarbilder sowie über die Änderung der Verlagerungsrichtung konvektiver Systeme ein, welche sogar bei Superzellen, die bisweilen gekrümmte Zugbahnen aufweisen können, in der Regel deutlich unter 30° innerhalb von 5 min liegt (z. B. Kunz et al., 2018).

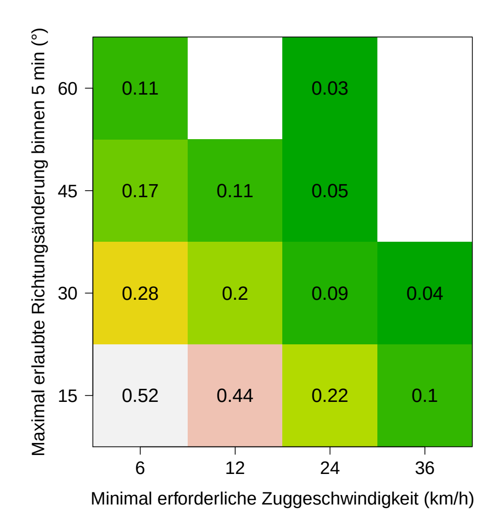
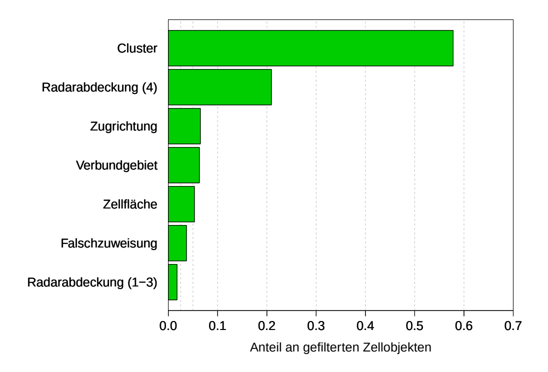


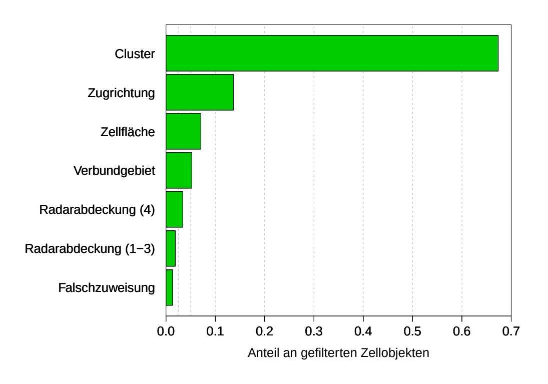
Abbildung 4.6: Ergebnis der Sensitivitätsuntersuchung zum Filter für die Verlagerungsrichtung für den Testzeitraum 27. Mai – 26. Juni 2016. Auf der Abszisse ist die zur Anwendung des Filters minimal notwendige Verlagerungsgeschwindigkeit $c_{Z,krit}$ aufgetragen. Die Ordinate gibt die innerhalb eines Zeitschritts von 5 min maximal erlaubte Änderung der geglätteten Verlagerungsrichtung $\alpha_{Z,krit}$ an. Die farblich unterlegten Zahlen stehen für den Anteil an Zellobjekten, die der Filter für die Verlagerungsrichtung unter Verwendung unterschiedlicher Schwellenwerte zusätzlich aussortiert.

Gefilterter Datensatz

Nach der Anwendung aller oben beschriebenen Filter ergibt sich ein stark reduzierter, aber sinnvoller Datensatz. Für die Sommerhalbjahre 2011 – 2016 werden von ursprünglich 165 572 Zellobjekten allein 62 009 Zellobjekte, die nur einmal von KONRAD registriert wurden, aufgrund einer zu kurzen Lebensdauer aussortiert (37,5 %). Von den verbleibenden Objekten werden weitere 65 010 (39,3 %) durch die übrigen Filter aussortiert, sodass schließlich noch 38 553 Objekte übrig bleiben (23,3 %).



(a) Sommerhalbjahre 2011 – 2016



(b) 27. Mai – 26. Juni 2016

Abbildung 4.7: Übersicht über die Anteile von Zellobjekten, die durch verschiedene, in Reihe geschaltete Filterkriterien aussortiert werden, für (a) die Sommerhalbjahre 2011 – 2016 sowie (b) den Testzeitraum 27. Mai – 26. Juni 2016. Zuvor wurden bereits die Zellobjekte aussortiert, die KONRAD nur einmal registrierte.

Der größte Anteil an diesen 38 553 gefilterten Zellobjekten ist auf den Clusterfilter zurückzuführen (56,5 %; Abbildung 4.7a). Im Testzeitraum 27. Mai – 26. Juni 2016, der zur Filterentwicklung diente, liegt er sogar bei 67,4 % (Abbildung 4.7b). In diesem Zeitraum war die Verfügbarkeit der

Radardaten deutlich besser, sodass nur 3,4 % der Objekte aufgrund fehlender Daten von drei oder mehr Radaren (Teilfilter 4 der Filterung hinsichtlich der Radarabdeckung) aussortiert werden. Dieser Filter hat im Zeitraum 2011 – 2016 hingegen einen Anteil von ca. 20,5 %. Der hohe Anteil des Clusterfilters kann zumindest als Indiz für das häufige Auftreten von Multizellen, MCS oder Gewitterlinien in Deutschland angesehen werden.
Von den restlichen Filtern tragen der Filter für die Verlagerungsrichtung (6,4 %), die Filterung am Rand des Verbundgebiets (6,2 %) sowie der Filter für die Zellfläche (5,2 %) zur Aussortierung von Zellobjekten am meisten bei. Auf die Filterung von Fehlzuweisungen sind 3,6 % und auf die Teilfilter 1 – 3 der Filterung hinsichtlich der Radarabdeckung (zusammengefasst) 1,7 % der aussortierten Zellobjekte zurückzuführen. Es sei angemerkt, dass die Filter innerhalb der Detektions-Schleife in Reihe geschaltet sind. Daher ist es möglich, dass Zellobjekte, die aufgrund eines bestimmten Filters aussortiert wurden, noch durch einen anderen Filter aussortiert worden wären, wenn es ersteren nicht gäbe. So sind beispielsweise neben dem Anteil des Clusterfilters die Anteile des Filters für die Zellfläche und des Filters für die Verlagerungsrichtung im gesamten Zeitraum kleiner als im Testzeitraum, da mehr Zellobjekte schon vorher aufgrund der Filterung hinsichtlich der Radarabdeckung aussortiert wurden. Die erwähnten Anteile sind folglich nicht als isolierte Einzelbeiträge der Filter zu verstehen, sondern stellen den Einfluss der Filter in Kombination dar.

4.3.3 Berechnung von Umgebungsvariablen aus den COSMO-Modelldaten

Die Assimilationsanalysen von COSMO-EU (vgl. Kapitel 4.2.2 und 4.2.3) beinhalten eine Vielzahl von atmosphärischen Variablen. Beispielsweise sind einzelne, für das *Nowcasting* konvektiver Zellen interessante Umgebungsvariablen abgespeichert, deren Berechnung in den Modulen der Datennachbehandlung (*Postprocessing*) von COSMO bereits implementiert ist, wie

z. B. die konvektiv verfügbare potentielle Energie (ML-CAPE), der *Showalter Index* (SI), der vertikal integrierte Wasserdampfgehalt IWV oder die bodennahe Feuchteflusskonvergenz (vgl. Kapitel 2.1.2, 2.4 und Anhang A). Des Weiteren ist es möglich, anhand der Analysedateien die stündliche Gesamtniederschlagssumme zu berechnen.
Viele interessante Umgebungsvariablen wurden bislang jedoch nicht in die Datennachbereitung von COSMO eingearbeitet und sind daher nicht in den Assimilationsanalysen verfügbar. Die Analysedateien können jedoch zur Initialisierung des COSMO-Modells verwendet werden. Außerdem ist es möglich, die Module der Datennachbehandlung um Routinen zur Berechnung weiterer Umgebungsvariablen zu ergänzen. Dies erlaubt es, nach der Initialisierung von COSMO mittels der Analysedateien die erweiterten Module der Datennachbehandlung dazu zu nutzen, weitere Umgebungsvariablen aus den Modellvariablen zum Initialisierungszeitpunkt zu berechnen[5] (Tabelle 4.3).
Zum bereits implementierten Aufstieg von ML-Luftpaketen wird mit einer ähnlichen Methodik die Berechnung für MU-Luftpakete hinzugefügt (vgl. Kapitel 2.1.2). Zur Berechnung des ML-HKN wird auf die analytische Formel nach Romps (2017) zurückgegriffen. Für mehrere der dynamischen und thermodynamischen Variablen in Tabelle 4.3 werden Berechnungen für verschiedene Höhenschichten bzw. -intervalle vorgenommen (z. B. für die SRH, die *Lapse Rate* und den mittleren Horizontalwind). Die Berechnung der *Lapse Rate* wird beispielsweise unter anderem aus der Temperaturdifferenz zwischen 0 und 1 500 m über Grund oder zwischen dem 800 und 600 hPa Druckniveau durch lineare Interpolation auf einem vertikal äquidistanten Hilfsgitter realisiert. Dieses Hilfsgitter findet unter anderem für die Berechnung der SRH Anwendung, die zur Bestimmung der geschätzten Verlagerungsgeschwindigkeit der Zellen den zwischen 0 und 6 km vertikal gemittelten Horizontalwind benötigt (vgl. Kapitel 2.2.3).

[5] Für die vorliegende Konfiguration von COSMO ist eine Parallelisierung mittels 27×18 MPI-Prozessoren empfehlenswert.

Tabelle 4.3: Zusammenfassung der atmosphärischen Umgebungsvariablen, deren Berechnung in die Module der Datennachbereitung von COSMO implementiert wurde und die zusätzlich zu bereits vorhandenen Ausgabegrößen zur Verfügung stehen (vgl. Kapitel 2 und Anhang A). Die Berechnung der mit einem Stern (*) versehenen Niveaus war für ein ML-Luftpaket bereits implementiert.

Beschreibung	Abkürzung	Indikator	Einheit	Dimension
Dynamische Größen				
Deep Layer Shear	DLS	Zellorganisation	$\mathrm{m\,s^{-1}}$	2D
Medium Layer Shear	MLS	Zellorganisation	$\mathrm{m\,s^{-1}}$	2D
Low Level Shear	LLS	Tornadorisiko	$\mathrm{m\,s^{-1}}$	2D
Sturm-relative Helizität	SRH	Zellorganisation	$\mathrm{m^2\,s^{-2}}$	2D
Mittlerer Horizontalwind	$\overline{\mathrm{U}}$	Zellverlagerung	$\mathrm{m\,s^{-1}}$	2D
Thermodynamische Größen				
Temperatur- und Feuchtemaße				
Pseudopotentielle Temperatur	θ_{ps}	Energiegehalt	K	3D
Taupunktstemperatur	τ	Feuchtegehalt	K	3D
Feuchttemperatur	T_F	Feuchtegehalt	K	3D
Mittlere rel. Feuchte über HKN	$\mathrm{RH_{HKN}}$	Feuchte mittl. Troposp.	%	2D
Lapse Rate	LR	Instabilität	$\mathrm{K\,m^{-1}}$	2D

(Fortsetzung Tabelle 4.3)

Beschreibung	Abkürzung	Indikator	Einheit	Dimension
Konvektive Indizes				
K-Index	—	Instabilität	K	2D
Total Totals	TT	Instabilität	K	2D
Vertical Totals	VT	bedingte Instabilität	K	2D
Lifted Index (ML)	LI	latente Instabilität	K	2D
Deep Convective Index	DCI	latente Instabilität	K	2D
KO-Index	—	potentielle Instabilität	K	2D
Vertikaldifferenz von θ_{ps}	$\Delta\theta_{ps}$	potentielle Instabilität	K	2D
Grenzhöhen				
Hebungskondensationsniveau*	HKN	Wolkenuntergrenze	m	2D
Niveau freier Konvektion*	NFK	potentielle Instabilität	m	2D
Niveau des neutralen Aufstiegs*	NNA	Zellmächtigkeit	m	2D
Kombinierte Parameter				
Supercell Composite Parameter	SCP	Superzellenwahrscheinl.	—	2D
Significant Hail Parameter	SHIP	Hagelwahrscheinlichkeit	—	2D
Bulk Richardson Number	BRN	Zellorganisation	—	2D

Die Ausgabe der 3D-Variablen (z. B. pseudopotentielle Temperatur θ_{ps}, Vertikalgeschwindigkeit ω) erfolgt anschließend als Viel-Flächen-Felder auf 16 Druckniveaus über das gesamte Modellgebiet von COSMO-EU, die der 2D-Variablen als Ein-Flächen-Feld (DWD-interne Bezeichnungen für 2D/3D-Feldstrukturen, GRIB2-Format; z. B. Schulz und Schättler, 2014). Eine Glättung der Felder über 9×9 horizontale Gitterpunkte reduziert lokale scharfe Gradienten der Umgebungsvariablen. Der so generierte umfangreiche Datensatz ist somit auch für zukünftige wissenschaftliche Studien, die ganz Europa betreffen, sehr nützlich[6].

Durch die Nachberechnung der weiteren Umgebungsvariablen für die Sommerhalbjahre 2011 – 2016 stehen insgesamt 83 Variablen in stündlicher Auflösung zur Verfügung. Von einer Erhöhung der stündlichen Auflösung wurde aufgrund eines größeren Zeit- und Rechenaufwands sowie eines zu hohen anfallenden Speicheraufwands abgesehen. Ohnehin ist die zeitliche und räumliche Verfügbarkeit einer so großen Anzahl von Umgebungsvariablen im Vergleich zu den meisten der weltweit bislang durchgeführten Studien bereits sehr hoch. Viele Studien der vergangenen Jahre aus Europa und den USA, die Gewitter-, Hagel- oder ähnliche konvektive Ereignisse anhand verschiedener Konvektionsparameter und Indizes beschreiben, basieren entweder auf grob aufgelösten Reanalysedaten (z. B. Kaltenböck et al., 2009; Brooks, 2009; Ukkonen et al., 2017; Westermayer et al., 2017) oder Radiosondenaufstiegen (z. B. Kunz, 2007; Mohr und Kunz, 2013; Púčik et al., 2015). Die COSMO-EU-Daten mit einer zeitlichen Auflösung von 1 h und einer horizontalen Auflösung von ca. 7 km sind damit eine sehr gute Datengrundlage (vgl. Kapitel 4.2.1; Miller und Mote, 2018).

[6] Gleichzeitig zu der Berechnung der Umgebungsvariablen im COSMO-EU-Setup erfolgte eine analoge Berechnung der Variablen für die Sommerhalbjahre 2011 – 2017 im COSMO-DE-Setup (vgl. Kapitel 4.2.1), welche ebenfalls in zukünftigen Untersuchungen Verwendung finden können.

Die Überprüfung der Implementierung der neuen Umgebungsvariablen erfolgte zum einen durch einen umfangreichen Vergleich der berechneten Werte für ausgewählte Zeitpunkte mit Werten, die aus Daten von Radiosondenaufstiegen berechnet wurden. Zum anderen konnte für einige Variablen ein Vergleich der berechneten Felder mit verschiedenen Literaturquellen oder frei zugänglichen Reanalysekarten vorgenommen werden. Zur Veranschaulichung einiger neu berechneter Umgebungsvariablen ist in Anhang D ein synoptisches Fallbeispiel, der 28. Juli 2013, betrachtet (Abbildungen D.1 und D.2).

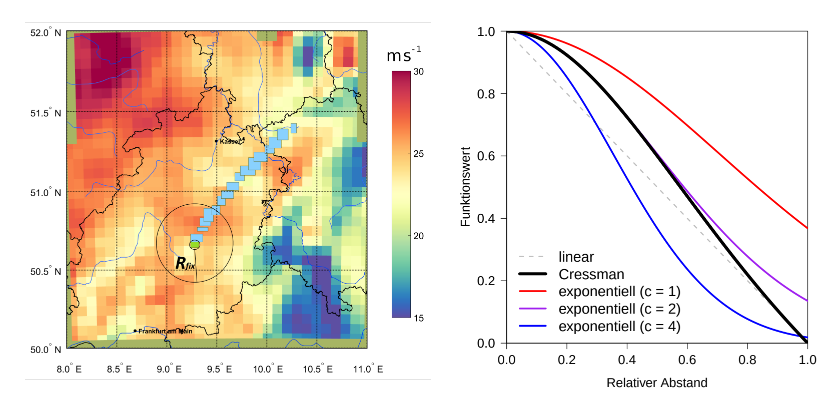
4.3.4 Zusammenführung der Zellverfolgungs- und Modelldaten

Nach der erfolgreichen Implementierung und zeitaufwändigen Berechnung der weiteren konvektionsrelevanten Umgebungsvariablen mittels COSMO-EU werden diese mit den aus KONRAD abgeleiteten und gefilterten Lebenszyklen der Zellobjekte (Kapitel 4.3.2) kombiniert. Die Zusammenführung der objektbezogenen Daten der Lebenszyklen und der gitterbasierten Daten der Umgebungsvariablen geschieht durch eine Erweiterung des Datensatzes der Lebenszyklen um repräsentative Werte der Umgebungsvariablen zu einem kombinierten objektbezogenen Datensatz. Ein Zellobjekt trägt dort zu jedem Zeitpunkt seines Lebenszyklus nicht nur die in den Tabellen 4.1 und 4.2 gelisteten Attribute, sondern zusätzlich verschiedene Informationen über die jeweils vorliegenden Umgebungsbedingungen.
Anstatt einem Zellobjekt zu jedem Zeitpunkt seines Lebenszyklus die Werte der Umgebungsvariablen des dem Zentroid am nächsten gelegenen Gitterpunkts in COSMO-EU zum nächstgelegenen Analysezeitpunkt zuzuschreiben, wird zunächst eine lineare Interpolation der COSMO-EU-Felder auf die zeitliche Auflösung der Zellobjekte (5 min) durchgeführt. Anschließend wird

zu jedem Detektionszeitpunkt eine Umgebung um das Zellobjekt gelegt, innerhalb derer der Algorithmus alle Gitterpunkte von COSMO-EU zur Berechnung repräsentativer Werte der Umgebungsvariablen für das registrierte Zellobjekt zum jeweiligen Detektionszeitpunkt berücksichtigt. In dieser Umgebung lassen sich verschiedene statistische Eigenschaften der Werte der Umgebungsvariablen bestimmen. Realisiert wird diese Idee im ersten Schritt durch die Konstruktion eines Kreises um das durch $\lambda_{Z,W}$, $\lambda_{Z,O}$, $\phi_{Z,S}$ und $\phi_{Z,N}$ aufgespannte Rechteck, das die Zellfläche einrahmt (vgl. Kapitel 4.1.2). Dieser Kreis schließt die vier Eckpunkte des Rechtecks ein und dessen Radius heiße im Folgenden Zellradius R_Z. Auch eine elliptische Form wäre möglich, um der horizontalen Anisotropie der Zellobjekte Rechnung zu tragen. Da die Zellflächen jedoch in derselben Größenordnung wie die horizontale Fläche einer Gitterbox von COSMO-EU liegen, ist die exakte geometrische Form unbedeutend. Im zweiten Schritt erfolgt eine Erweiterung des Zellradius R_Z um einen festen Wert R_{fix}, sodass der Gesamtradius der dadurch definierten Umgebung einer Zelle durch $R_U = R_Z + R_{fix}$ gegeben ist (s. u.; Abbildung 4.8a). Diese adaptive Umgebung mit dem Radius R_U ist demnach für größere Zellobjekte größer als für kleinere.

Im Fall, dass bestimmte Umgebungsvariablen wie beispielsweise das NFK oder der SHIP nicht an allen Gitterpunkten innerhalb von R_U vorliegen, fordert ein Kriterium, dass an mindestens $N_{GP,min}$ Gitterpunkten, welche gleichzeitig einem Anteil von mindestens $f_{GP,min}$ aller Gitterpunkte innerhalb R_U entsprechen müssen, Werte vorliegen müssen. Wenn nur an wenigen Gitterpunkten innerhalb von R_U Werte vorliegen, sind diese nicht unbedingt repräsentativ für die Umgebung einer Zelle. Ist dieses Kriterium jedoch erfüllt, erfolgt die Zuweisung mehrerer statistischer Maße der Umgebungsvariablen zum jeweiligen Zellobjekt (s. u.). Für das abstandsgewichtete Mittel bestimmen sich die Gewichte W_i mittels der von Cressman (1959) definierten Formel

$$W_i = \frac{R_U^2 - r_i^2}{R_U^2 + r_i^2} \,. \tag{4.12}$$



(a) Zur Illustration der Zellumgebung　　**(b)** Verschiedene Gewichtungsfunktionen

Abbildung 4.8: (a) Kombinierte Darstellung der DLS ($\mathrm{m\,s^{-1}}$), berechnet mit COSMO-EU, und der Zugbahn einer langlebigen Gewitterzelle über Nordhessen am 11. September 2011 (15 UTC), einem Tag, an dem mehrere Superzellen mit großem Hagel über die Mitte und den Norden Deutschlands zogen (z. B. Fluck, 2018). Das KONRAD-Zellobjekt ist in fünfminütlichen Abständen in Form von hellblauen Rechtecken eingezeichnet, die alle ihm zugehörigen Radarpixel einrahmen. Die erste Zelldetektion (grünes Rechteck) war um 15 UTC. Um das Zellobjekt herum sind zwei schwarze Kreise zur Illustration des Zellradius R_Z, des Gesamtradius R_U sowie von R_{fix} eingezeichnet. (b) Vergleich verschiedener Gewichtungsfunktionen für den relativen Abstand r_r vom Zentroid. Die Exponentialfunktionen haben die Form $\exp(-cr_r^2)$.

Darin bezeichnet r_i den Abstand r zwischen dem i-ten Gitterpunkt und dem Zentroid. Mit der Definition des relativen Abstands vom Zentroid in Bezug auf den Gesamtradius $r_r = rR_U^{-1}$ folgt, dass für $r_r \to 0$ die Cressman-Funktion sehr ähnlich wie eine Exponentialfunktion verläuft, während für $r_r \to 1$ der Verlauf eher linear wird, sodass sie bei $r = R_U$ den Funktionswert 0 erreicht (Abbildung 4.8b). Anschließend folgt eine Normierung der Gewichte, sodass sie in Summe 1 ergeben. Darüber hinaus speichert der Algorithmus die Information über die Anzahl der Gitterpunkte innerhalb von R_U sowie den Wert von R_U für jedes Zellobjekt ab.

Drei Aspekte dieser Vorgehensweise für die Zuweisung der Umgebungsvariablen zu den Lebenszyklen der Zellobjekte in Bezug auf die Eigenschaften der vorliegenden Daten werden in der folgenden Zusammenstellung kurz diskutiert:

(a) Einzelne Umgebungsvariablen variieren auf Zeitskalen deutlich unterhalb von einer Stunde (z. B. bodennahe Feuchteflusskonvergenz), sodass eine zeitliche Interpolation der einstündigen COSMO-EU-Werte auf 5 min nicht zwangsweise realistischere Werte liefert. Die räumliche Mittelung führt zudem zur Glättung kleinskaliger Variationen in den Werten der Variablen. Statistische Analysen mit solchen Variablen müssen daher, sollte eine Interpolation stattfinden, mit besonderer Vorsicht interpretiert werden.

(b) Der Kreis um die Zellobjekte erfasst die durch die zeitliche lineare Interpolation auftretende Varianz der Umgebungswerte. Dies ist ein weiteres Argument dafür, nicht nur den nächstgelegenen Gitterpunkt, sondern eine größere Anzahl von nahegelegenen Gitterpunkten zur Zuweisung der Umgebungsbedingungen zu wählen.

(c) Zur Charakterisierung der präkonvektiven Bedingungen wäre die bevorzugte Berücksichtigung von Gitterpunkten stromabwärts einer Zelle eine intuitive Festlegung, insbesondere im Bereich von Frontalzonen. Da die Datenassimilation von COSMO-EU bis zum 3. September 2014 jedoch kein *Latent Heat Nudging* durchführte und Radardaten daher nicht assimilierte (vgl. Kapitel 4.2.1), „kennen“ die vor diesem Datum berechneten Analysen des Modells die beobachtete Position der beobachteten konvektiven Zellen nicht. Die nach diesem Datum erstellten Analysen sind näher an der beobachteten Niederschlagsverteilung. Um unabhängig von der Anwendung des *Latent Heat Nudgings* zu untersuchen, ob sich der analysierte Niederschlag auf die Werte der Umgebungsvariablen innerhalb der über R_U definierten Zellumgebungen

auswirkt, erfolgt die Berechnung des Datensatzes auf zwei verschiedene Arten: einmal wie oben beschrieben und einmal mit einer zusätzlichen Filterung bezüglich des analysierten Niederschlags (s. u.). Für Letztere gehen nur Umgebungswerte von denjenigen Gitterpunkten in die Bestimmung der Umgebungsbedingungen ein, an denen innerhalb der vorangegangenen Stunde weniger als 1 mm Niederschlag analysiert wurden. Mit diesem Filter werden vor allem die Gitterpunkte aussortiert, an denen durch die Konvektionsparametrisierung von COSMO-EU (oder durch das *Latent Heat Nudging*) die simulierten Umgebungsbedingungen potentiell modifiziert wurden, welche damit als nicht repräsentativ für die präkonvektiven Bedingungen vermutet werden.

Damit sind mehrere Randbedingungen für die Zuweisung der Umgebungsvariablen zu den Lebenszyklen der Zellobjekte frei wählbar:

(1) Wie groß soll R_{fix} gewählt werden?

(2) Welcher Zeitpunkt der Modelldaten ist am besten geeignet?

(3) Ist eine zeitliche Interpolation der Modelldaten sinnvoll?

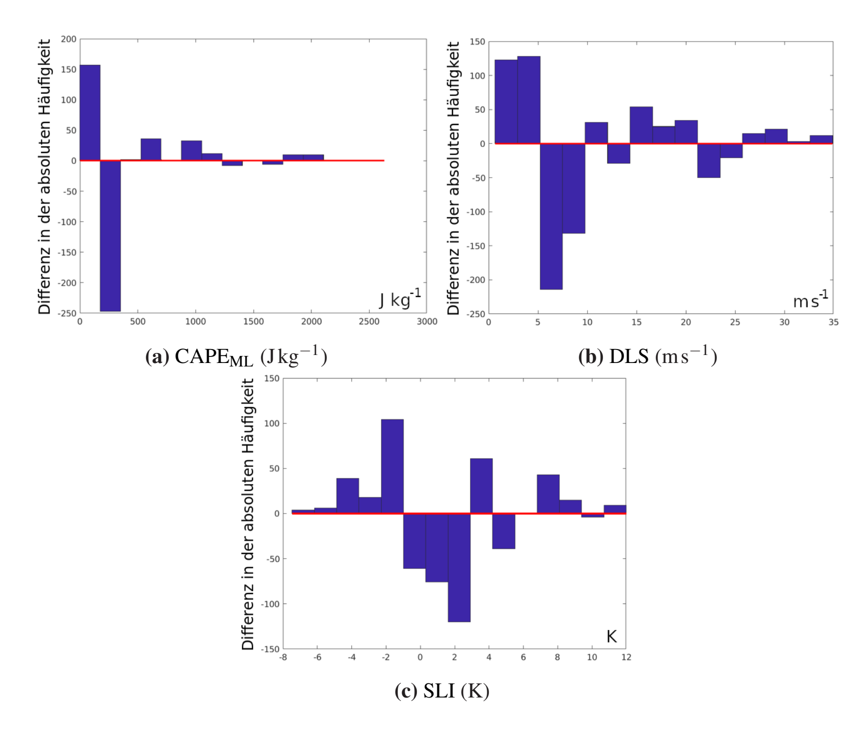
(4) Wie viele Gitterpunkte charakterisieren die Umgebungsbedingungen am besten?

(5) Bewirkt eine Filterung bezüglich des modellierten Niederschlags einen Mehrwert (s. o. Diskussionspunkt (c))? Und falls ja, welcher Schwellenwert für die stündliche Niederschlagssumme ist am sinnvollsten zu wählen?

(6) Welche statistischen Maße soll der Algorithmus für weitere Untersuchungen abspeichern und was muss bei deren Berechnung genau beachtet werden?

Zur Beantwortung dieser Fragen wurden mehrere Untersuchungen der Sensitivität bezüglich verschiedener Randbedingungen durchgeführt:

(1) **Umgebungsradius** R_U: Variation von R_{fix} zwischen 20 und 50 km
Die meisten konvektionsrelevanten Variablen variieren auf horizontalen Skalen deutlich unter 100 km. Haklander und van Delden (2003) fanden die beste Korrelation zwischen aus Radiosondendaten abgeleiteten Umgebungsvariablen und detektierten Blitzen für einen Radius von 20 km um die zugehörige Station. Hamann et al. (2019) verwenden für das *Nowcasting*-System COALITION von MeteoSchweiz (vgl. Kapitel 2.4) für jeden ihrer Prädiktoren einen festen Umgebungsradius von 11,5 km. Zöbisch et al. (2020) legten um ihre Zellobjekte für Umgebungsvariablen aus hochaufgelösten COSMO-DE-Vorhersagen eine Umgebung mit einem größeren Radius von 50 km fest, um den Effekt von simulierten konvektiven Zellen auf die Modellfelder möglichst gering zu halten. R_U darf in der vorliegenden Arbeit für Umgebungsvariablen aus COSMO-EU-Assimilationsanalysen nicht zu klein, aber auch nicht zu groß gewählt werden. Je kleiner R_U ist, desto weniger Gitterpunkte gehen ein und desto größer ist der Einfluss von lokalen Variabilitäten, welche durch die bei der Berechnung der Umgebungsvariablen durchgeführte Glättung jedoch nicht allzu groß sein sollten (vgl. Kapitel 4.3.3). Je größer R_U ist, desto größer ist die Gefahr, dass auch Gitterpunkte jenseits von Frontalzonen oder Luftmassengrenzen zur Berechnung der statistischen Maße beitragen. Ein beispielhafter Vergleich der Häufigkeiten der Werte der Umgebungsvariablen zwischen einem Umgebungsradius R_U mit $R_{fix} = 25$ km und einem mit $R_{fix} = 50$ km deutet stark darauf hin, dass die Unterschiede nicht allzu groß sind (Abbildung 4.9). Bei einer horizontalen Auflösung von 7 km liegen für $R_{fix} = 25$ km in knapp 94 % der Fälle 45 – 66 COSMO-EU-Gitterpunkte innerhalb von R_U, was als ausreichend für die Beschreibung der Umgebung beurteilt wird. Daher wird $R_{fix} = 25$ km festgelegt.



(a) $\mathrm{CAPE_{ML}}$ ($\mathrm{J\,kg^{-1}}$) **(b)** DLS ($\mathrm{m\,s^{-1}}$)

(c) SLI (K)

Abbildung 4.9: Differenzen der absoluten Häufigkeit der Werte verschiedener Umgebungsvariablen (arithmetischer Mittelwert der Umgebung) für den Testzeitraum 27. Mai – 26. Juni 2016 zwischen einem Umgebungsradius R_U mit $R_{fix} = 25$ km und einem mit $R_{fix} = 50$ km. Insgesamt gehen in den Vergleich 14 891 Werte ein, also die Werte zu allen Detektionszeitpunkten der 3 749 Zellobjekte, die die Filterung aus Kapitel 4.3.2 passiert haben.

(2) **Zeitpunkt der Zuweisung**: t_D; $t_D - 30$ min; $t_D - 60$ min mit Detektionszeitpunkt t_D

Einige konvektionsrelevante Variablen variieren auf zeitlichen Skalen unter 1 h. Für die meisten der betrachteten Variablen sind die Unterschiede zwischen den (gemittelten) Umgebungswerten zum Zeitpunkt t_D und den Werten 30 oder 60 min vorher jedoch vernachlässigbar klein (nicht gezeigt). Lediglich für wenige Variablen wie z. B. die bodennahe relative Feuchte, die LLS

oder verschiedene *Lapse Rates* sind bei weniger als 5 % der Zellobjekte deutlich höhere/niedrigere Werte vorzufinden. Daher fällt die Wahl des Zeitpunkts für die Zusammenführung der Daten auf t_D.

(3) **Zeitliche lineare Interpolation**: ja oder nein
Da viele Umgebungsvariablen auf deutlich größeren, räumlichen Skalen als der Umgebung eines Zellobjekts R_U variieren, stellt sich heraus, dass eine zeitliche lineare Interpolation etwas genauere Werte für die Umgebungsvariablen zu den jeweiligen Detektionszeitpunkten bestimmen kann, ohne dass diese stark von möglichen kleinskaligen (unbekannten) Variationen in der Realität abweichen (s. o. Diskussionspunkt (a); Zöbisch et al., 2020).

(4) **Mindestanzahl von Gitterpunkten**: 15 bis 50
Der maximale Gesamtradius R_U liegt für knapp 94 % der Zellobjekte mit $R_{fix} = 25$ km zwischen 27 und 32 km (der Rest liegt darüber), da für die meisten Zellobjekte $R_Z \ll R_U$ gilt. Aufgrund der Anzahl von meist $45-66$ COSMO-EU-Gitterpunkten innerhalb von R_U wird die Mindestanzahl von Gitterpunkten innerhalb von R_U auf $N_{GP,min} = 30$ festgelegt. Gleichzeitig muss die Anzahl einem Anteil von mindestens $f_{GP,min} = 50\,\%$ aller Gitterpunkte innerhalb R_U entsprechen, damit von einer ausreichenden Repräsentativität ausgegangen werden kann.

(5) **Schwellenwert für die Niederschlagsfilterung**: 0,5; 1,0; 5,0 mm
Je kleiner der Schwellenwert ist, desto mehr Gitterpunkte werden vom Algorithmus bei der Zuweisung der Umgebungswerte zu den Zellobjekten als Gitterpunkte mit signifikantem Niederschlag erkannt, die als nicht repräsentativ für die präkonvektiven Umgebungsbedingungen angenommen werden und daher nicht in die Berechnungen für die Zuweisung eingehen. Ein hoher Schwellenwert dagegen macht die Filterung vernachlässigbar. Damit weisen beispielsweise nur 2 % der Umgebungen aller betrachteten Zellobjekte bei einem Schwellenwert von 5 mm so viele Gitterpunkte mit signifikantem Niederschlag auf, dass Werte von $N_{GP,min} < 30$ oder $f_{GP,min} < 50\,\%$ auftreten.

Mit einem Schwellenwert von 1 mm ergibt sich ein Anteil von rund 10 %. Somit führt die Niederschlagsfilterung in Kombination mit der Forderung einer Mindestanzahl von Gitterpunkten in der Zellumgebung zwangsläufig zu einer geringeren Anzahl von repräsentativen Umgebungen. Ohne Niederschlagsfilterung ist der kombinierte Datensatz folglich größer. Da qualitativ nur geringe Unterschiede in der Verteilung der Umgebungsvariablen der verbleibenden Zellobjekte zu erkennen sind, findet für die weiteren Untersuchungen ab Kapitel 5 nur der Datensatz ohne Niederschlagsfilterung Anwendung.

(6) **Statistische Maße**

Als Kriterium für eine plausible Zuweisungsmethode wird neben der physikalischen Sinnhaftigkeit und einem geringen Rechenaufwand auch ein konsistenter zeitlicher Verlauf der zugewiesenen statistischen Maße der Umgebungsvariablen gefordert, d. h. die Werte der Maße sollten möglichst keine signifikanten Sprünge innerhalb eines (fünfminütigen) Zeitschritts aufweisen. Dies ist für ein abstandsgewichtetes Mittel, das über Gleichung (4.12) bestimmt wird, und das arithmetische Mittel sowie deren entsprechende Standardabweichungen der Fall. Dasselbe gilt zudem für verschiedene Perzentilwerte (0, 25, 50, 75, 100). Diese zwei Mittelwerte und Standardabweichungen ergeben mit den fünf Perzentilwerten insgesamt neun statistische Maße, welche die Verteilung der Werte der Umgebungsvariablen charakterisieren. Der Algorithmus speichert diese für jedes Zellobjekt zu jedem Detektionszeitpunkt ab. Eine eigens entwickelte Methode zur Zuweisung, die auf realistischere Werte bei einer bi- oder trimodalen Verteilung der Umgebungswerte innerhalb von R_U abzielt, findet aufgrund von weniger konsistenten zeitlichen Verläufen keine Anwendung. Meist ist die Variation der so erhaltenen Werte der neun ausgewählten statistischen Maße während des Lebenszyklus nur gering (s. Kapitel 5.2.1).

5 Lebenszyklus und Umgebungsbedingungen konvektiver Zellen

Eine Analyse des in Kapitel 4.3.4 beschriebenen kombinierten Datensatzes aus den zusammengestellten Lebenszyklen konvektiver Zellen und den vorherrschenden Umgebungsbedingungen kann das Verständnis der Entwicklung konvektiver Zellen verbessern, welches für das *Nowcasting* von Gewittern nützlich ist. Hierfür ist zunächst eine Analyse der Lebenszyklen ohne die Berücksichtigung der Umgebungsvariablen hilfreich (Kapitel 5.1). Daneben bietet eine Analyse der aus Modelldaten gewonnenen Umgebungsbedingungen die Möglichkeit, die verschiedenen Umgebungsvariablen und deren Korrelationen untereinander besser einzuordnen (Kapitel 5.2). Durch die Kombination der Lebenszyklen und Umgebungsvariablen gelingt es anschließend aufzuzeigen, dass einige Umgebungsvariablen und Zellattribute konvektive Zellen hinsichtlich unterschiedlicher Charakteristika zu einem gewissen Maß unterscheiden können (Kapitel 5.3).

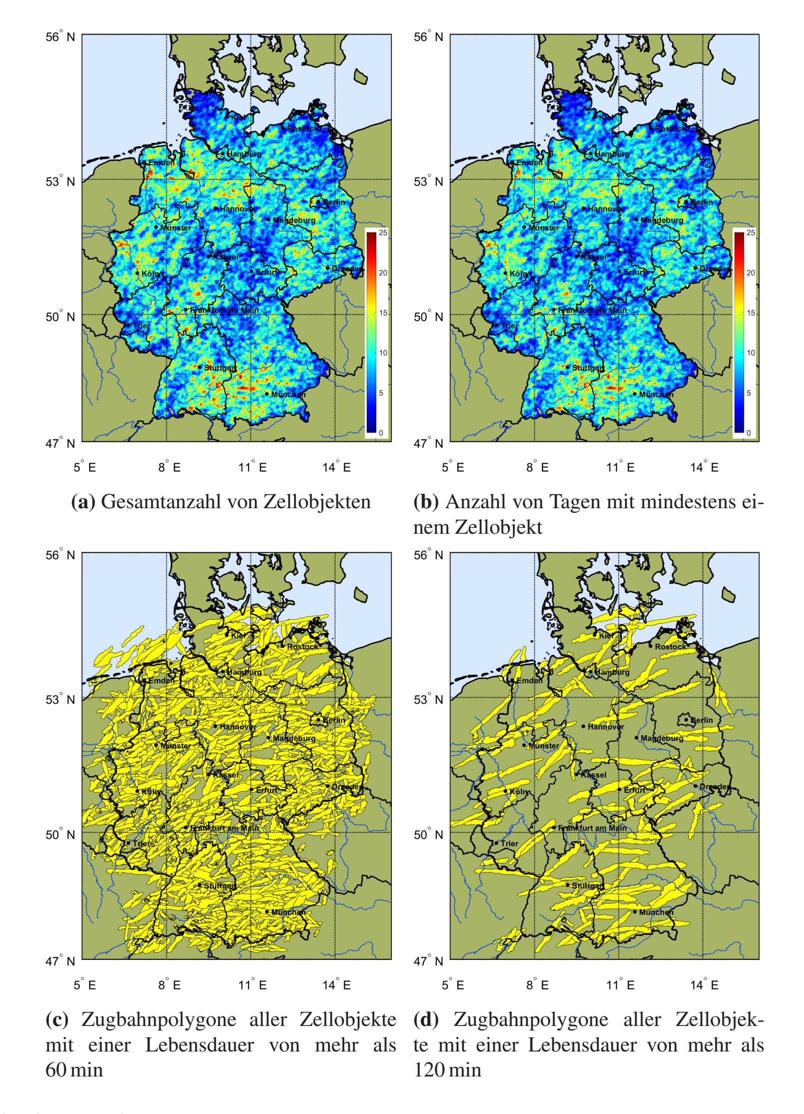
5.1 Statistische Analyse der Zellobjekte

5.1.1 Merkmale der Zellattribute

Die 38 553 gefilterten Zellobjekte aus den Sommerhalbjahren 2011 – 2016 verteilen sich über das gesamte Gebiet der Bundesrepublik Deutschland (Abbildung 5.1). Die Zugbahnen der konvektiven Zellen können durch Polygone, die aus den Zellobjekten konstruiert sind, näherungsweise dargestellt

werden (vgl. Schmidberger, 2018). Die Zusammenstellung der Polygone geschieht dergestalt, dass zu jedem Detektionszeitpunkt vom Mittelpunkt des einrahmenden Rechtecks rechtwinklig zur Verlagerungsrichtung eine horizontale Ausdehnung angenommen wird, die der Länge der Diagonalen D des einrahmenden Rechtecks entspricht (Abbildung 5.2; vgl. Kapitel 4.1.2). Der Anfang (das Ende) der Polygone ergibt sich durch eine konstruierte Verlängerung der detektierten Zugbahn um die Hälfte der Länge der Diagonalen gegen (in) die Verlagerungsrichtung. Den Karten in den Abbildungen 5.1a+b liegt ein $1 \times 1\,\mathrm{km}^2$-Gitter zugrunde. Die absolute Häufigkeit für jeden Gitterpunkt entspricht der Anzahl von Zugbahnpolygonen, die diesen Gitterpunkt einschließen.

Regionen mit einer Häufung von identifizierten Zellobjekten sind die Schwäbische Alb, das Alpenvorland, ein Gebiet von der Rhein-Neckar-Region bis zur Wetterau und dem Vogelsberg sowie Nordrhein-Westfalen und Niedersachsen. Hier traten lokal insgesamt etwa 25 Zellen auf, was etwa vier Zellen pro Jahr entspricht. Die Anzahl von Tagen mit mindestens einem Zellobjekt liegt meist nur wenig unter der beobachteten Objektanzahl. Piper und Kunz (2017) bestimmten basierend auf Blitzdaten für Deutschland abhängig von der Region eine mittlere Anzahl von Gewittertagen (mindestens 5 Blitze innerhalb einer Gitterzelle mit einer Ausdehnung von $10 \times 10\,\mathrm{km}^2$) zwischen meist 5 und 15. Die in Abbildung 5.1b dargestellten Werte sind im Vergleich dazu niedriger und ergeben sich zum einen dadurch, dass die Auflösung des hier verwendeten Gitters deutlich höher gewählt wurde. Zum anderen spielt die Filterung des hier vorliegenden Datensatzes eine Rolle, aufgrund derer nur lediglich 23,3 % aller (bzw. 37,2 % aller mindestens zweimal registrierten) Zellobjekte vertreten sind. Gerade räumlich ausgedehnte Multizellen und MCS sind aufgrund des Clusterfilters meist herausgefiltert worden (vgl. Kapitel 4.3.2). Die ersten drei der oben genannten Regionen (Schwäbische Alb, Alpenvorland, Rhein-Main-Gebiet und Umgebung) sind bekannt für das häufige Auftreten von Gewittern (z. B. Wapler und James, 2015; Piper und Kunz, 2017;



(a) Gesamtanzahl von Zellobjekten

(b) Anzahl von Tagen mit mindestens einem Zellobjekt

(c) Zugbahnpolygone aller Zellobjekte mit einer Lebensdauer von mehr als 60 min

(d) Zugbahnpolygone aller Zellobjekte mit einer Lebensdauer von mehr als 120 min

Abbildung 5.1: Räumliche Verteilung der Zellobjekte im Untersuchungszeitraum der Sommerhalbjahre 2011 – 2016. Zu Darstellungszwecken sind in (a) und (b) die Häufigkeiten für Gitterpunkte über dem Meer und außerhalb Deutschlands ausgeblendet.

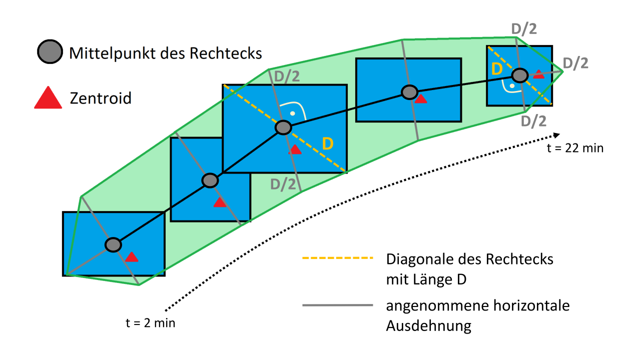
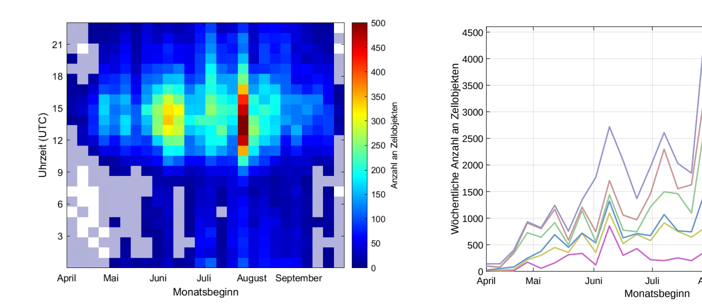


Abbildung 5.2: Schematische Darstellung der Konstruktion der mit den Zellobjekten assoziierten Zugbahnpolygone am Beispiel eines Zellobjekts mit der Lebensdauer $T_Z = 22$ min.

Taszarek et al., 2019). Die vergleichsweise hohe Anzahl von Zugbahnen in Nordwestdeutschland, wo die Anzahl von Gewittertagen pro Jahr im Mittel deutlich geringer als in der Südhälfte ist, lässt sich auf die konvektiv sehr aktive Periode im Mai und Juni 2016 zurückführen (s. u.; vgl. Piper et al., 2016). Zellobjekte mit einer langen Lebensdauer von mehr als 60 min (1 096 Objekte bzw. 2,8 % von 38 553) und 120 min (121 Objekte bzw. 0,3 %) sind ebenfalls in ganz Deutschland zu beobachten[1]. Die Zugbahn des Objekts mit der höchsten Lebensdauer (257 min) gehört zu einem Gewitter, das sich am 25. August 2012 gegen 13 UTC nordöstlich von Hannover bildete und dessen Zugbahn sich bis kurz vor die polnische Grenze erstreckte. Unter den zehn Objekten mit der längsten Lebensdauer finden sich mehrere durch Augenzeugenberichte bestätigte Superzellen, während einige prominente Beispiele

[1] Die Lebensdauer eines Zellobjekts wird im Folgenden der KONRAD-internen Zuweisung entsprechend stets mit 7, 12, 17 min etc. angegeben (vgl. Kapitel 4.1.2). Häufig findet im Folgenden auch eine Beschreibung eines Zellattributs durch einen Bezug auf den Zeitpunkt der ersten Detektion statt. Beispielsweise sind die Bezeichnungen „Zellfläche 10 min nach der ersten Detektion", „Zellfläche zum Zeitpunkt der dritten Detektion" und „Zellfläche im Alter von 12 min" äquivalent.



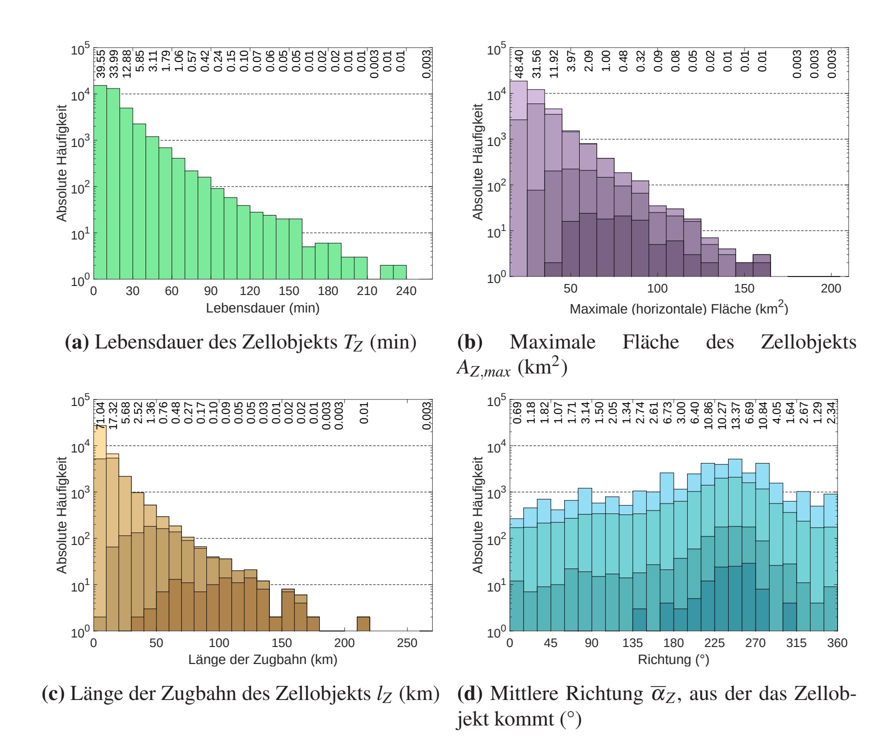
(a) Häufigkeitsverteilung der Zellobjekte nach Datum und Uhrzeit (UTC)

(b) Wöchentliche Häufigkeiten der Zellobjekte

Abbildung 5.3: (a) Absolute Häufigkeitsverteilung der Zellobjekte je nach Datum und Uhrzeit (UTC) des jeweiligen Tags. Die Uhrzeit entspricht dem Zeitpunkt der ersten Detektion. Die Klassifikation erfolgt auf Wochen- bzw. Stundenbasis. Klassen mit weniger als zehn Objekten sind transparent dargestellt. (b) Wöchentliche Häufigkeiten der Zellobjekte für die Sommerhalbjahre 2011 – 2016. Die Linien zeigen jeweils die Summe der Objekte aller Sommerhalbjahre von 2011 bis einschließlich zum angegebenen Jahr.

wie die Superzelle vom 11. September 2011 (Rheinland-Pfalz bis Brandenburg; z. B. Fluck, 2018) oder die Superzellen vom 27. und 28. Juli 2013 (in der Region um Wolfsburg bzw. Reutlingen; vgl. Kunz et al., 2018) nicht im Datensatz enthalten sind.

Im Untersuchungszeitraum wurde isolierte Konvektion vornehmlich zwischen 10 und 19 UTC, also zwischen 12 und 21 Uhr MESZ, beobachtet (Abbildung 5.3a). Besonders im Hoch- und Spätsommer kam es jedoch auch zu nächtlicher Konvektion. Diese Verteilung hat große Ähnlichkeiten mit den für verschiedene Teile Mitteleuropas mit Hilfe von Blitzdaten bestimmten Verteilungen von Schulz et al. (2005), Novák und Kyznarová (2011), Wapler (2013) und Piper und Kunz (2017). Insbesondere die letzte Juliwoche ist in Abbildung 5.3a auffällig. Hier entwickelten sich in den Jahren 2013, 2014 und 2016 besonders viele konvektive Zellen (Abbildung 5.3b). Das nachmittägliche Maximum der Häufigkeit Anfang Juni lässt sich vor allem auf die konvektiv aktiven Perioden der Jahre 2011 sowie 2016 zurückführen.



(a) Lebensdauer des Zellobjekts T_Z (min)

(b) Maximale Fläche des Zellobjekts $A_{Z,max}$ (km^2)

(c) Länge der Zugbahn des Zellobjekts l_Z (km)

(d) Mittlere Richtung $\overline{\alpha}_Z$, aus der das Zellobjekt kommt (°)

Abbildung 5.4: Absolute Häufigkeitsverteilung verschiedener Zellattribute in logarithmischer Darstellung für alle 38 553 Zellobjekte. Die am Oberrand angegebenen Zahlen geben entsprechend die relative Häufigkeit (%) an. Die Abbildungen (b)–(d) sind zusätzlich um die Verteilungen aller Zellobjekte mit einer Lebensdauer von mehr als 15, 60 bzw. 120 min ergänzt. Je dunkler der Farbton, desto höher der Schwellenwert. In (d) entspricht 0° Norden, 90° Osten, 180° Süden und 270° Westen.

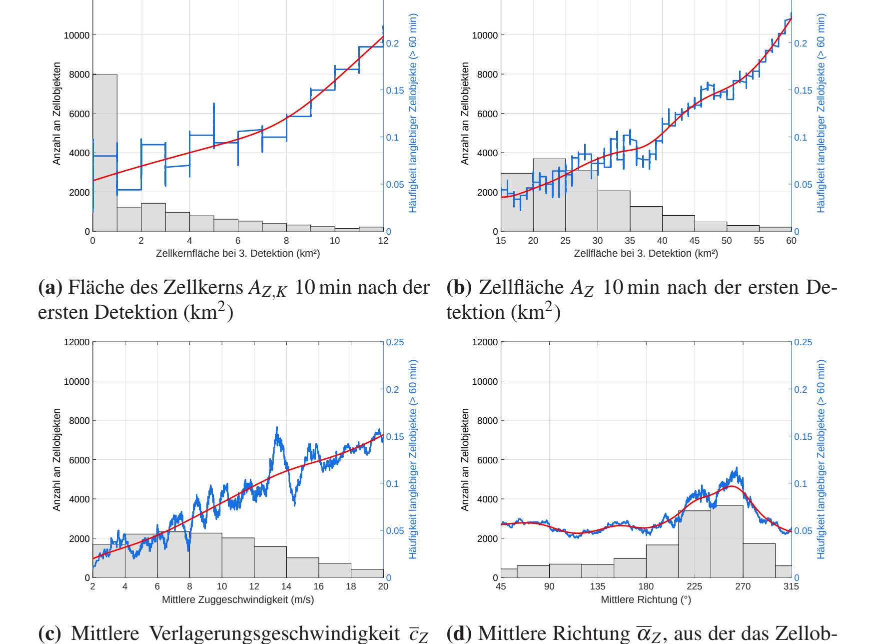
Den größten Anteil der 38 553 Zellobjekte stellen erwartungsgemäß konvektive Zellen mit einer kürzeren Lebensdauer dar, wobei sowohl das Zellverfolgungsverfahren von KONRAD als auch die hier angewendete Filterung einen Einfluss auf die Verteilung haben. Daher ergibt sich eine sehr schiefe Verteilungsfunktion hinsichtlich der Lebensdauer (Abbildung 5.4a), die ein zentrales Grundproblem für die Entwicklung statistischer Modelle zur Vorhersage der Lebensdauer darstellt (vgl. Kapitel 2.4; Davini et al., 2012;

Wapler, 2021). Ähnlich schiefe Verteilungen ergeben sich für die maximale Fläche der Zellobjekte (Fläche mit einem Reflektivitätsfaktor von $Z \geq 46$ dBZ; vgl. Tabelle 4.1) während ihrer Lebenszyklen (Abbildung 5.4b) sowie die Zugbahnlänge (Abbildung 5.4c). Dabei stellen die Zellobjekte mit einer langen Lebensdauer von mehr als 60 min einen großen Teil des rechten Rands der Gesamtverteilung der maximalen Zellfläche und der Zugbahnlänge, wobei von den Intervallen ganz rechts wiederum Zellobjekte mit einer sehr langen Lebensdauer von mehr als 120 min einen großen Teil stellen. Dennoch unterscheiden sich Zellen mit einer (sehr) langen Lebensdauer untereinander teilweise deutlich in ihrer maximalen Fläche und der Länge ihrer Zugbahn (und damit ihrer Verlagerungsgeschwindigkeit). Die Zugbahn des Zellobjekts mit der längsten Lebensdauer (25. August 2012) war beispielsweise 210 km lang, die maximale Fläche betrug jedoch lediglich 72 km^2. Eine Superzelle am 27. April 2015 zog hingegen deutlich langsamer, was zu einer Zugbahnlänge von 100 km führte, jedoch mit einer maximalen Fläche von 157 km^2.

Ein großer Anteil aller Zellobjekte kommt aus dem Sektor 195°−285°, zog also etwa von (süd-)westlichen in (nord-)östliche Richtungen (Abbildung 5.4d). Da das Zentroid der Zellobjekte auf den Daten des gerasterten Radarkomposits beruht, erkennt man besonders für die Zellobjekte mit einer kurzen Lebensdauer von weniger als 15 min, dass bestimmte Richtungen aufgrund der Gittergeometrie bevorzugt auftreten (z. B. knapp 6 000 Objekte, die in einem kleinen Bereich von 0,02 rad $\approx$ 1,14° um 90, 180, 270 und 360° liegen). Dieser Gittereffekt wird mit steigender Lebensdauer geringer, zumal der Algorithmus eine Glättung der Richtungswinkel über drei Detektionszeitpunkte vornimmt (vgl. Kapitel 4.3.2). Beispielsweise liegen im oben genannten Sektor 61,4 % aller Zellobjekte mit einer Lebensdauer von mehr als 15 min. Damit herrscht eine sehr hohe Übereinstimmung der hier vorliegenden Verteilung mit den Ergebnissen von Wapler und James (2015), Wapler (2021) und Schmidberger (2018). Erstere Studien basieren ebenfalls auf Stichproben von KONRAD-Daten, letztere auf 9 798 Hagelzugbahnen

aus dem Zellverfolgungsalgorithmus TRACE-3D. Berücksichtigt man nur die Verteilung der Zellobjekte mit langer Lebensdauer, steigt der Anteil der beobachteten Objekte im Sektor $195° - 285°$ auf 72,1 % (Lebensdauer länger als 60 min) bzw. 85,1 % (Lebensdauer länger als 120 min; vgl. Abbildung 5.1d). Eine solche südwestliche Anströmung transportiert häufig feucht-warme Luftmassen aus dem südwest-europäischen Raum nach Mitteleuropa, welche die Gewitterbildung in Deutschland begünstigen (Kapsch et al., 2012; Piper und Kunz, 2017; Mohr et al., 2019).

Im nächsten Schritt wird die relative Häufigkeit von Zellobjekten mit langer Lebensdauer in Abhängigkeit von verschiedenen Zellattributen betrachtet (Abbildung 5.5). Um sinnvolle Werte für die betrachteten Zellattribute zu erhalten, gehen in diese Analyse nur Zellobjekte ein, die mindestens viermal registriert wurden ($T_Z > 15$ min). Die relative Häufigkeit für einen festen Abszissenwert (blaue Linie) bestimmt sich hier, indem zunächst die 500 (Abbildungen 5.5a−c) bzw. 1 000 (Abbildung 5.5d) Zellobjekte ausgewählt werden, deren entsprechender Wert diesem Abszissenwert am nächsten liegt. Innerhalb dieser Auswahl wird anschließend der Anteil von Zellen mit einer langen Lebensdauer ($T_Z > 60$ min) bestimmt.
Bei einer schnellen Intensivierung bzw. einem schnellen horizontalen Wachstum zu Beginn des Lebenszyklus steigt die Wahrscheinlichkeit für langlebige Zellen. Intensiviert sich eine Zelle binnen 10 min nach der ersten Detektion, sodass die Fläche des Zellkerns des assoziierten Zellobjekts $A_{Z,K}$ (Teilbereich der Zellfläche mit einem Reflektivitätsfaktor von $Z \geq 55$ dBZ; vgl. Tabelle 4.1) auf etwa $A_{Z,K} > 8 - 10\,\text{km}^2$ angewachsen ist, verdreifacht sich etwa die Wahrscheinlichkeit für eine lange Lebensdauer von mehr als 60 min gegenüber Zellobjekten mit $A_{Z,K} = 0\,\text{km}^2$ (Abbildung 5.5a). Gleiches gilt, wenn ein Zellobjekt in der gleichen Zeitspanne auf eine Zellfläche A_Z von mehr als etwa $A_Z = 45\,\text{km}^2$ anwächst (Abbildung 5.5b). Ursache hierfür könnte eine vorhergehende schnelle Intensivierung des Aufwindbereichs der Zellen sein, welche sowohl das vertikale Wachstum als auch die horizontale



(a) Fläche des Zellkerns $A_{Z,K}$ 10 min nach der ersten Detektion (km^2)

(b) Zellfläche A_Z 10 min nach der ersten Detektion (km^2)

(c) Mittlere Verlagerungsgeschwindigkeit $\overline{c}_Z$ des Zellobjekts (m s^{-1})

(d) Mittlere Richtung $\overline{\alpha}_Z$, aus der das Zellobjekt kommt (°)

Abbildung 5.5: Absolute Häufigkeit verschiedener Zellattribute für alle Zellobjekte mit einer Lebensdauer von mehr als 15 min ($T_Z > 15$ min; Histogramm, linke Ordinate). Außerdem: relative Häufigkeit verschiedener Zellattribute für Zellobjekte mit einer langen Lebensdauer ($T_Z > 60$ min) bezüglich der Zellobjekte mit $T_Z > 15$ min (blaue Kurve, rechte Ordinate). In (a) und (b) existieren aufgrund der diskreten Werteverteilung der Zellattribute für einen Abszissenwert viele Ordinatenwerte. Zusätzlich dargestellt ist eine Glättung der relativen Häufigkeiten mittels lokaler linearer Regression (*Local Linear Kernel Regression*; vgl. Cleveland, 1979; rote Kurve). Details siehe Fließtext.

Ausbreitung der hochreichenden Konvektionszelle fördert. Dadurch kommt es zu rascher Niederschlagsbildung in einem ausgedehnten Luftvolumen, welche sich kurze Zeit später in hohen Werten des Reflektivitätsfaktors niederschlägt. Eine ähnliche schnelle Zellentwicklung wurde beispielsweise bei der bekannten Superzelle vom 28. Juli 2013 beobachtet (Kunz et al., 2018).

Bei niedrigen über den Lebenszyklus gemittelten Verlagerungsgeschwindigkeiten von $\overline{c}_Z \leq 5\,\mathrm{m\,s^{-1}}$, die meist mit einer gradientschwachen synoptisch-skaligen Hintergrundströmung und damit verbunden mit einer geringen vertikalen Windscherung einhergehen, ist die Wahrscheinlichkeit für eine lange Lebensdauer vier- bis fünfmal geringer als bei $\overline{c}_Z > 15\,\mathrm{m\,s^{-1}}$ (Abbildung 5.5c). Dies kann mit den bevorzugten Organisationsformen der Ereignisse erklärt werden, da unter windschwachen Bedingungen vornehmlich isolierte Einzelzellen auftreten, deren Aufwindbereich infolge des selbst produzierten Niederschlags frühzeitig abgebaut wird (vgl. Kapitel 2.2). Bei südwestlicher bis westlicher Anströmung ist die Wahrscheinlichkeit, dass ein Zellobjekt eine Lebensdauer von mehr als 60 min erreicht, knapp doppelt so hoch wie bei anderen Strömungsverhältnissen (Abbildung 5.5d). Darüber hinaus sind etwa 89,8 % aller Zellobjekte mit Zugbahnen länger als 50 km im Sektor Süd-Südwest bis West-Nordwest zu finden, was sowohl auf eine längere Lebensdauer als auch auf schnellere Verlagerungsgeschwindigkeiten zurückzuführen ist (nicht gezeigt).

5.1.2 Beschreibung des Lebenszyklus der Zellobjekte

Bereits aufgrund der unterschiedlichen Organisationsformen konvektiver Zellen (Kapitel 2.2) wird deutlich, dass eine allgemeingültige Definition des Lebenszyklus für alle Formen nicht möglich ist. Dennoch zielen viele Untersuchungen darauf ab, auf Basis einzelner Fallstudien oder Statistiken einer Stichprobe möglichst verallgemeinernde Aussagen treffen zu können, um dieses Verständnis in (operationellen) *Nowcasting*-Verfahren – am besten mittels einer einfachen, aber universellen Methode – anzuwenden (vgl. Kapitel 2.4).

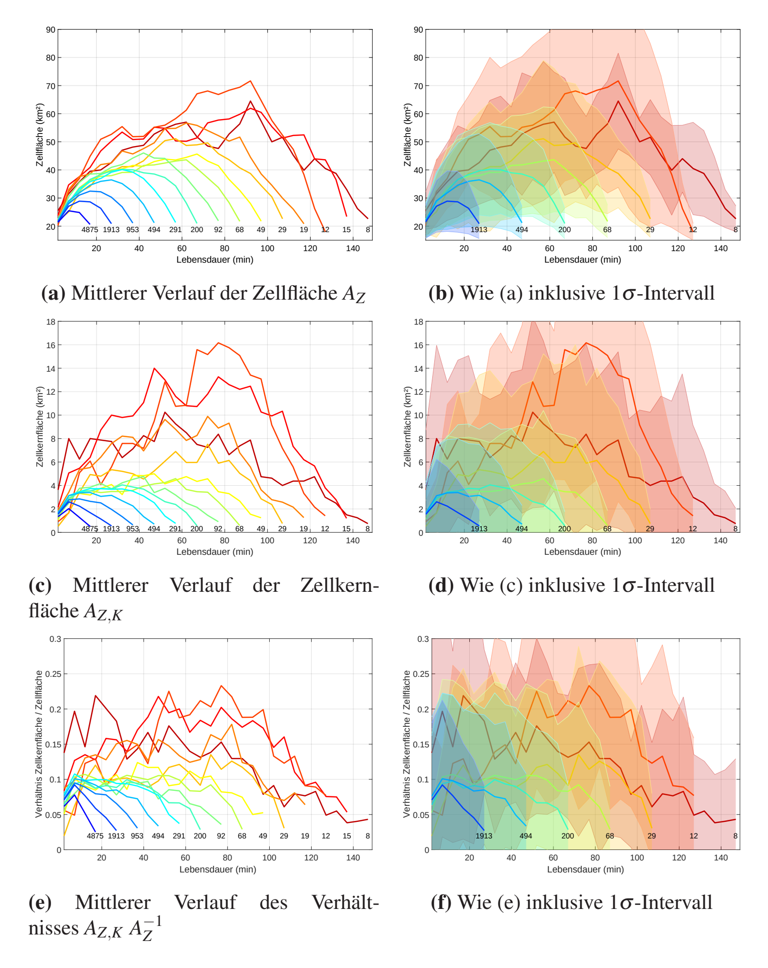
Parabelansatz

Ältere Untersuchungen von KONRAD-Daten zeigen, dass die mittlere Entwicklung der Zellflächen in etwa die Form einer nach unten geöffneten Parabel oder einer halben Periode einer Sinusfunktion

hat (Blahak et al., 2018; Wapler, 2021), weshalb der DWD zurzeit einen Parabelansatz testet[2] (vgl. Kapitel 4.1.2; Feger et al., 2019; Werner, 2020). Auch Weusthoff und Hauf (2008) beobachteten bereits, dass sich die Fläche (sowie die über die Fläche gemittelte Regenrate) von postfrontalen Einzelzellen auf diese Weise entwickelt. Davini et al. (2012) berichteten, dass konvektive Zellen ihre maximale Intensität bereits in der ersten Hälfte ihres Lebenszyklus erreichen, während die maximale Zellfläche erst in der zweiten Hälfte angenommen wird. Neuere Untersuchungen von Brisson et al. (2018) legen nahe, dass die höchste mittlere Regenrate bereits etwa nach einem Viertel der Lebensdauer auftritt. Im Folgenden wird in Anlehnung an diese Untersuchungen zunächst ein Parabelansatz formuliert und diskutiert. In Kapitel 5.3.1 wird dieser um die Berücksichtigung einer Umgebungsvariablen erweitert und in Kapitel 6.3.2 für den Zweck einer quantitativen Modellstudie zur Abschätzung der Lebensdauer konvektiver Zellen verwendet.

Der vorliegende Datensatz bestätigt den mittleren parabelförmigen Verlauf der Fläche der KONRAD-Zellobjekte A_Z (Abbildung 5.6a). Die maximale beobachtete Zellfläche steigt mit zunehmender Lebensdauer an. Allerdings ist die Variabilität der einzelnen Entwicklungen der Zellobjekte sehr hoch, sodass sich die Variationsbereiche (1σ-Intervall) der Objektgruppen unterschiedlicher Lebensdauer stark überlappen (Abbildung 5.6b). Hinzu kommt, dass sich insbesondere in den ersten $15-30$ min des Lebenszyklus beispielsweise die mittleren Flächenentwicklungen der Objekte mit einer Lebensdauer von $T_Z \in (45\,\text{min}\,;\,90\,\text{min})$ kaum von denen mit einer noch längeren Lebensdauer unterscheiden. Dies lässt die Schlussfolgerung zu, dass sich alleine auf dieser Basis eine (deterministische) Abschätzung der zu erwartenden Lebensdauer von konvektiven Zellen, insbesondere innerhalb der ersten halben Stunde, als eher schwierig erweisen wird (Details folgen in Kapitel 6.3.2 und Anhang C).

[2] Verwendung als internes Lebenszyklusmodell für KONRAD3D in Kombination mit einem Ensemble Kalman Filter (EnKF).



(a) Mittlerer Verlauf der Zellfläche A_Z

(b) Wie (a) inklusive 1σ-Intervall

(c) Mittlerer Verlauf der Zellkernfläche $A_{Z,K}$

(d) Wie (c) inklusive 1σ-Intervall

(e) Mittlerer Verlauf des Verhältnisses $A_{Z,K}\,A_Z^{-1}$

(f) Wie (e) inklusive 1σ-Intervall

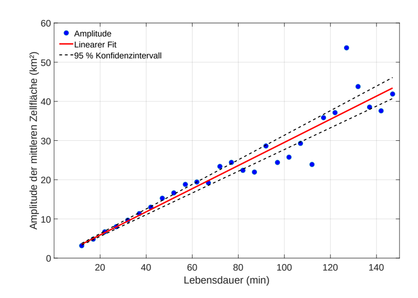
Abbildung 5.6: Mittlere zeitliche Entwicklung (a) der Fläche der Zellobjekte, (c) der Zellkernfläche sowie (e) deren Verhältnis, sortiert nach der Lebensdauer der Objekte. Unterschiedliche Linienfarben indizieren unterschiedliche Werte für die Lebensdauer. Die Zahlen an den Enden der Linien geben die Anzahl von Objekten an, die zur Mittelung beigetragen haben. Nur jede zweite Linie der fünfminütlich aufgelösten Zellstatistik ist der Übersicht halber eingezeichnet. (b), (d) und (f): Wie (a), (c) und (e), nur dass nur jede vierte Linie eingezeichnet ist, dafür aber mit einem der Standardabweichung entsprechenden Variationsbereich (1σ-Intervall).

Etwas differenzierter stellt sich die Entwicklung der Fläche des Zellkerns $A_{Z,K}$ bzw. des Verhältnisses der Fläche des Zellkerns zur gesamten Zellfläche dar (Abbildungen 5.6c−f). Dieses Verhältnis kann als Proxy für die mittlere Regenrate oder die Intensität des Zellobjekts angesehen werden. Ist dieses Verhältnis groß, ist ein großer Teil des Zellobjekts auf Regionen mit einem sehr hohen Reflektivitätsfaktor von $Z \geq 55\,\mathrm{dBZ}$ zurückzuführen. Doch auch große Zellobjekte mit kleinem Verhältnis können intensive Zellkerne beinhalten. Das mittlere Verhältnis von Zellkernfläche zu Zellfläche liegt für Zellobjekte mit einer längeren Lebensdauer die meiste Zeit des Lebenszyklus – und auch schon sehr frühzeitig – oberhalb von $A_{Z,K}\ A_Z^{-1} = 0{,}1$. Objekte mit einer kürzeren Lebensdauer hingegen erreichen das maximale Verhältnis bereits kurz nach der ersten Detektion mit etwa $A_{Z,K}\ A_Z^{-1} = 0{,}1$, welches anschließend im Mittel abfällt (Abbildung 5.6e; vgl. Brisson et al., 2018). Auch hieraus lässt sich folgern, dass ein rasches, intensives Wachstum zu Beginn des Lebenszyklus ein Indikator für eine lange Lebensdauer sein kann (vgl. Davini et al., 2012).

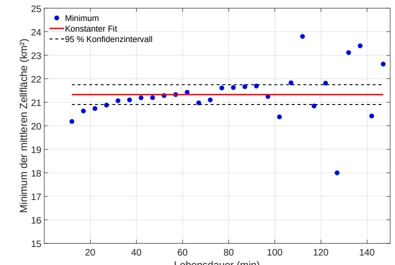
Die Darstellung des Parabelansatzes zur Beschreibung der zeitlichen Entwicklung der mittleren Zellfläche durch eine Funktionenschar mit dem Scharparameter T_Z lautet (analog zu Weusthoff und Hauf, 2008):

$$A_Z^{(T_Z)}(t) = A_{Z,min}^{(T_Z)} + \mathscr{A}^{(T_Z)} - \frac{\mathscr{A}^{(T_Z)}}{(T_Z/2)^2}\left(t - \frac{T_Z}{2}\right)^2 . \tag{5.1}$$

Darin ist $\mathscr{A}^{(T_Z)} = A_{Z,max}^{(T_Z)} - A_{Z,min}^{(T_Z)}$ die Amplitude der Entwicklung der jeweiligen mittleren Zellflächen. Der Korrelationskoeffizient auf Basis der 38 553 Zellobjekte für die Korrelation zwischen der Amplitude und der Lebensdauer T_Z liegt bei einem hohen Wert von $r_P \approx 0{,}74$ ($r_S \approx 0{,}73$), d. h. die Abhängigkeit der Amplitude von der Lebensdauer kann näherungsweise als linear angenommen werden, sodass $\mathscr{A}^{(T_Z)} \approx c_A T_Z$ mit



(a) Bestimmung von $\mathscr{A}(T_Z)$. Der Regressionskoeffizient ist $c_A = 0{,}295\,\text{km}^2\,\text{min}^{-1}$, der *RMSE* liegt bei rund $4{,}2\,\text{km}^2$.



(b) Bestimmung von $A_{Z,min}(T_Z)$. Der Regressionskoeffizient ist $\mu_A = 21{,}326\,\text{km}^2$, der *RMSE* liegt bei rund $1{,}1\,\text{km}^2$.

Abbildung 5.7: Überblick über (a) die Amplituden und (b) die Minima der Verläufe der mittleren Zellflächen in Abhängigkeit von der Lebensdauer. Der lineare Fit in (a) erfolgt ohne konstanten Term. Außerdem gehen nur Werte für die Lebensdauer zwischen 10 und 150 min ein, um wenige Zellobjekte mit einer langen Lebensdauer von mehr als 150 min nicht zu stark zu gewichten.

dem mittels linearer Regression (vgl. Kapitel 3.3.1) bestimmten Koeffizienten $c_A = 0{,}295\,\text{km}^2\,\text{min}^{-1}$ gilt (Abbildung 5.7a). Das Minimum der Zellfläche wird zum Zeitpunkt der ersten Detektion angenommen, welche aufgrund des Kriteriums zur Detektion einer Zelle in KONRAD für die meisten Zellobjekte eine ähnliche Größe hat (vgl. Kapitel 4.1.2; Wapler, 2021). Über einen konstanten Fit erhält man daher $A_{Z,min}^{(T_Z)} \approx \mu_A$, wobei gilt: $\mu_A = 21{,}326\,\text{km}^2$ (Abbildung 5.7b). Damit wird Gleichung (5.1) zu:

$$A_Z^{(T_Z)}(t) = \mu_A + c_A T_Z - \frac{c_A T_Z}{(T_Z/2)^2}\left(t - \frac{T_Z}{2}\right)^2 = \mu_A + 4 c_A t \left(1 - \frac{t}{T_Z}\right) \; . \qquad (5.2)$$

Die so erhaltene Parabelschar (Abbildung 5.8a) ist in ihrem Wertebereich nach oben hin durch $A_{Z,krit}(t) = \mu_A + 4 c_A t$ limitiert. Wie man diesen Ansatz zur Abschätzung der zu erwartenden Lebensdauer oder maximalen Zellfläche anwenden kann, beschreiben die Kapitel 6.3.2 und 6.4.2.

Gleichung (5.1) lässt sich durch wenige mathematische Operationen weiter vereinfachen und man erhält

$$\frac{A_Z^{(T_Z)}(t) - A_{Z,min}^{(T_Z)}}{\mathscr{A}^{(T_Z)}} = 1 - 4\left(\frac{t}{T_Z} - \frac{1}{2}\right)^2$$
$$\Longleftrightarrow A_Z^*(t^*) = 4t^*\,(1 - t^*) \tag{5.3}$$

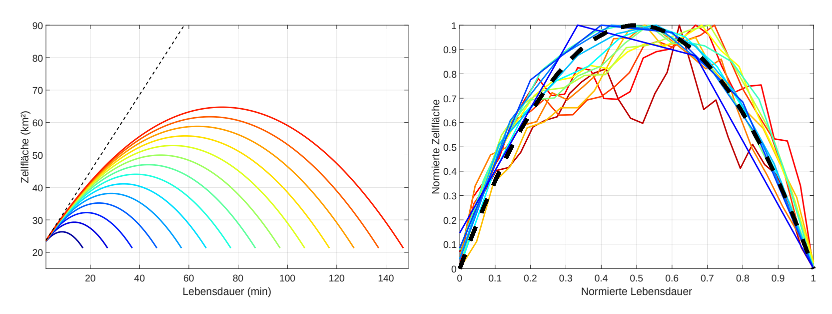
mit den Normierungsvorschriften

$$A_Z^* = \frac{A_Z^{(T_Z)}(t) - A_{Z,min}^{(T_Z)}}{\mathscr{A}^{(T_Z)}} \qquad ; \qquad t^* = \frac{t}{T_Z}\ . \tag{5.4}$$

Durch diese Normierung reduziert sich die Parabelschar auf eine einzige Parabel, die den mittleren Verlauf der mit der Amplitude normierten Zellfläche während des Lebenszyklus aller Zellobjekte beschreibt (Abbildung 5.8b; vgl. Weusthoff und Hauf, 2008). Betrachtet man die normierten Verläufe in Abhängigkeit von der Lebensdauer, ist zu erkennen, dass sich mit steigender (absoluter) Lebensdauer das Maximum der Zellfläche zu einem späteren (relativen) Zeitpunkt des Lebenszyklus verschiebt (ca. zwei Drittel des Lebenszyklus; vgl. Davini et al., 2012). Eine mögliche Erklärung könnte sein, dass Zellen mit einem besonders intensiven und breiten Aufwindbereich und damit verbunden einer großen vertikalen Erstreckung eine lange Lebensdauer erreichen. Diese dehnen sich während des Lebenszyklus aufgrund der Begrenzung durch die Tropopause zunehmend horizontal aus (vgl. Kapitel 2.2). Möglicherweise erreichen diese Zellen die größte Zellfläche erst nach dem Zeitpunkt der höchsten maximalen Intensität.

Strömungsfeldansatz

Eine alternative Betrachtungsweise der zeitlichen Entwicklung der Zellfläche zum Parabelansatz ist die Darstellung mit Hilfe eines Strömungsfelds im Zustandsraum $\mathscr{Z}$, dessen Dimensionen das Zellalter und die Zellfläche



(a) Parabelschar gemäß Gleichung (5.2)

(b) Normierter Verlauf der Zellfläche sowie Parabel gemäß Gleichung (5.3)

Abbildung 5.8: (a) Parabelschar des analytischen Modells gemäß Gleichung (5.2) für die zeitliche Entwicklung der Zellfläche $A_Z(t)$ mit der Lebensdauer T_Z als Scharparameter. Die Limitierung des Wertebereichs ist durch die schwarz gestrichelte Linie gekennzeichnet. (b) Mittlerer Verlauf der normierten Zellfläche A_Z^* bezüglich des jeweiligen Lebenszyklus. Schwarz gestrichelt ist die Kurve des normierten Parabelmodells aus Gleichung (5.3).

aufspannen (Abbildung 5.9). Die Beschreibung des Strömungsfelds

$$\mathbf{v}_Z(t,A_Z) = v_t(t,A_Z)\mathbf{e}_t + v_A(t,A_Z)\mathbf{e}_A \tag{5.5}$$

mit den Einheitsvektoren $\mathbf{e}$ und den entsprechenden Komponenten v ist beispielsweise durch den Median $\langle\cdot\rangle$ der beobachteten lokalen (Euler'schen) Tendenzen der beiden Zellattribute Zellfläche (A_Z) und Zellalter (t) bezüglich der Zeit (t') innerhalb eines bestimmten Teilgebiets von $\mathscr{Z}$ möglich:

$$v_t(t,A_Z) = \left\langle \frac{\partial t}{\partial t'} \right\rangle (t,A_Z) = 1 \ , \tag{5.6}$$

$$v_A(t,A_Z) = \left\langle \frac{\partial A_Z}{\partial t'} \right\rangle (t,A_Z) \ . \tag{5.7}$$

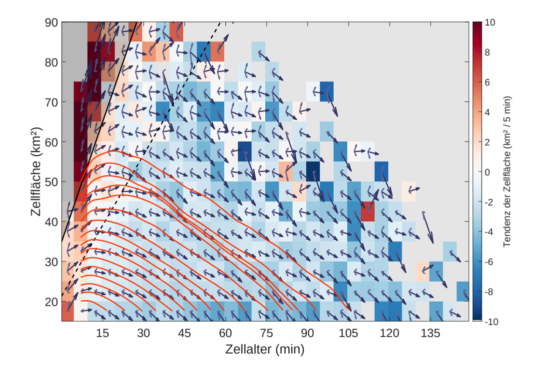
Die Stromlinien $\mathbf{s}_v$ dieses Strömungsfelds, die überall tangential zum lokalen Strömungsvektor liegen, charakterisieren folglich mögliche Entwicklungen der Zellfläche. Numerisch erfolgt die Bestimmung der Stromlinien über die

Stromliniengleichung (Stromlinienparameter ξ)

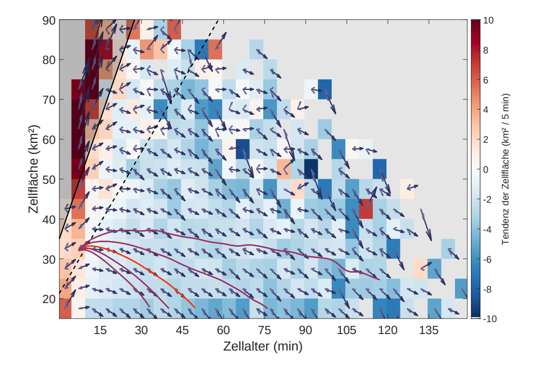
$$\frac{\mathrm{d}\mathbf{s}_\nu}{\mathrm{d}\xi} = \mathbf{v}_Z\left(\mathbf{s}_\nu(\xi)\right) \tag{5.8}$$

beispielsweise mit Hilfe von Finite-Differenzen-Verfahren.
Zur Bestimmung des in Abbildung 5.9 dargestellten Strömungsfelds gehen alle Zellobjekte ein, die mindestens dreimal registriert wurden. Finden Zellobjekte mit noch kürzerer Lebensdauer Berücksichtigung, so weist das Strömungsfeld schon zu Beginn negative Tendenzen der Zellfläche auf. Die Tendenzen bestimmen sich über finite Differenzen: zu Beginn und am Ende des Lebenszyklus über einseitige Differenzen zweiter Ordnung und dazwischen über zentrierte Differenzen zweiter oder, sofern möglich, vierter Ordnung. Dadurch erfolgt zudem eine leichte Glättung der jeweiligen Tendenzverläufe der Zellobjekte. Nur solche Teilgebiete von $\mathscr{X}$, die mindestens zehn Zellobjekte aufweisen können, tragen zum Strömungsfeld bei. Für alle anderen wird angenommen, dass die wenigen beobachteten Tendenzen nicht repräsentativ genug sind.

Der Median der Tendenz der Zellfläche ist in den ersten 15 min nach dem Zeitpunkt der ersten Detektion meist positiv. Je größer die Zellfläche zum Zeitpunkt der zweiten oder dritten Detektion ist, desto größer ist die Tendenz für ein weiteres (schnelles) Wachstum. Schon nach 20 min ist der Median der Tendenz der Zellfläche meist negativ, wobei die Variabilität auch in dieser Analyse sehr hoch ist. Die negative Tendenz resultiert dabei aus der schiefen Verteilung des Datensatzes bezüglich der Lebensdauer (vgl. Abbildung 5.4a). Der Verlauf der Stromlinien in Abbildung 5.9a ist somit gestreckter als die Verläufe im Parabelansatz. Man sieht zudem eindrücklich, dass der oben hergeleitete Parabelansatz aus Gleichung (5.2) ein sehr schnelles Zellwachstum zu Beginn des Lebenszyklus ausschließt (Region links oberhalb der gestrichelten Linie).



(a) Strömungsfeld und Stromlinien, welche die Entwicklung von Zellobjekten unterschiedlich starken anfänglichen Wachstums charakterisieren.



(b) Strömungsfeld und Stromlinien, welche die Entwicklung von Zellobjekten gleich starken anfänglichen Wachstums charakterisieren, die sich jedoch im Anschluss wie das 40., 45., 50. (rot), 55. bzw. 60. Perzentil entwickeln.

Abbildung 5.9: Mittlere Entwicklungstendenzen der Zellfläche aller Zellobjekte eines bestimmten Alters innerhalb eines Bereichs der Zellfläche von $5\,\mathrm{km}^2$ (farbige Boxen). Darüber das Strömungsfeld, dargestellt durch entsprechende Bewegungsvektoren im Zellalter-Zellfläche-Raum (Median [dunkelblaue Pfeile] sowie (a) 25. und 75. Perzentil, (b) 40. und 60. Perzentil [hellere Pfeile, gekürzt]). Limitierungen des Parabelmodells $A_{Z,krit}(t) = \mu_A + 4c_A t$ sind als gestrichelte ($\mu_A = 21{,}326\,\mathrm{km}^2$, $c_A = 0{,}295\,\mathrm{km}^2\,\mathrm{min}^{-1}$) und durchgezogene ($\mu_A = 35\,\mathrm{km}^2$, $c_A = 0{,}5\,\mathrm{km}^2\,\mathrm{min}^{-1}$) Linien dargestellt. Rötliche Linien stellen Stromlinien dar. Der Strömungsvektor in einem Gebiet ist nur dort dargestellt, wo mindestens zehn Zellobjekte vorliegen.

Wapler (2021) konstatierte auf der Basis eines ähnlichen Datensatzes und einer ähnlichen Darstellung der Verläufe der Zellfläche als Parabel, dass die Wachstumsrate aller Zellobjekte zu Beginn recht ähnlich sei. Ein schnelles Wachstum der Zellfläche wird jedoch recht häufig beobachtet (s. o.) und durch das Strömungsfeld besser repräsentiert. Die sehr großen Tendenzen links oberhalb der durchgezogenen Linie, die $A_{Z,krit}(t)$ für sehr hohe Werte der Regressionsparameter $\mu_A = 35\,\mathrm{km}^2$ und $c_A = 0{,}5\,\mathrm{km}^2\,\mathrm{min}^{-1}$ darstellt (s. Kapitel 5.3.1), erfordern aufgrund der Berechnungsmethodik eine vorsichtige Interpretation. Trotz der umfangreichen Filterung sind einzelne fehlerhafte Lebenszyklen von Zellobjekten im Datensatz vorhanden, die den Filter für die Zellfläche (vgl. Kapitel 4.3.2) gerade so passiert haben. Während Zellobjekte mit einer Lebensdauer von mehr als 30 min und (zugleich) einer maximalen Zellfläche von mehr als $60\,\mathrm{km}^2$ zum Zeitpunkt der zweiten Detektion einen Median von etwa $30{-}35\,\mathrm{km}^2$ aufweisen, erreichen Zellobjekte mit $T_Z < 30\,\mathrm{min}$ und $A_{Z,max} > 60\,\mathrm{km}^2$ für diesen Zeitpunkt im Median eine Zellfläche von $57\,\mathrm{km}^2$ (nicht gezeigt). Da diese extrem schnelle Zellflächenentwicklung bis zum zweiten Detektionszeitpunkt kaum mit einer neu entstandenen konvektiven Zelle erklärbar ist, sind letztere eher als fehlerhafte Lebenszyklen einzuordnen, wobei sie ca. 13,4 % (141) aller Zellobjekte mit einer maximalen Zellfläche von mehr als $60\,\mathrm{km}^2$ stellen (1 052).

Die Wahl des Medians zur Festlegung des Strömungsfelds $\mathbf{v}_Z$ in Gleichung (5.7) ist zwar intuitiv und plausibel, jedoch keineswegs zwingend. Wählt man statt des Medians nur leicht variierte Perzentile, so ergeben sich bereits sehr unterschiedliche Verläufe der Stromlinien mit demselben Startpunkt im Zustandsraum $\mathscr{Z}$ (Abbildung 5.9b). Erst durch die Stromlinien wird die Auswirkung der Variabilität der Tendenzen auf den Verlauf der Zellfläche und auf die Lebensdauer deutlich. Eine Anwendung des Strömungsfelds zur Abschätzung der zu erwartenden Lebensdauer oder maximalen Zellfläche erfolgt ebenso wie für den Parabelansatz in den Kapiteln 6.3.2 und 6.4.2.

5.2 Analyse der Umgebungsbedingungen

Basierend auf dem kombinierten Datensatz (vgl. Kapitel 4.3.4) werden im Folgenden die atmosphärischen Umgebungsbedingungen der Zellobjekte genauer untersucht. Um eine bessere Lesbarkeit zu gewährleisten, werden Abkürzungen für die in Tabelle E.1 genauer beschriebenen und hier relevanten Umgebungsvariablen eingeführt (vgl. hierzu auch die Beschreibung der Variablen in Kapitel 2, Tabelle 4.3 bzw. Anhang A). Die Auswahl dieser Variablen erfolgt in Anlehnung an die noch folgenden Analysen aus Kapitel 5.3.1, die unter anderem das Unterscheidungsvermögen der Umgebungsvariablen hinsichtlich verschiedener Werte für die Lebensdauer der Zellobjekte untersuchen. Dadurch wird eine objektive Reduzierung der Variablenanzahl von 747 auf 33 Variablen erreicht, welche Redundanzen verringert. Wie in Kapitel 4.3.4 beschrieben sind für jede Variable verschiedene statistische Maße für die Zellumgebung verfügbar, welche in den Abbildungen mit einem Kürzel gemäß der Spalte „Statistik“ in Tabelle E.1 hinter dem Variablennamen charakterisiert, im Fließtext jedoch ausgespart werden[3].

5.2.1 Statistische Merkmale der Umgebungsvariablen

Im Folgenden werden exemplarisch einige Umgebungsvariablen diskutiert, die Maße für die thermische Stabilität, den Feuchtegehalt oder die vertikale Windscherung darstellen. In Bezug auf die Stabilität weisen knapp 73 % der Zellobjekte gemittelt über ihren Lebenszyklus eine $\mathrm{CAPE_{MU}}$ von weniger als

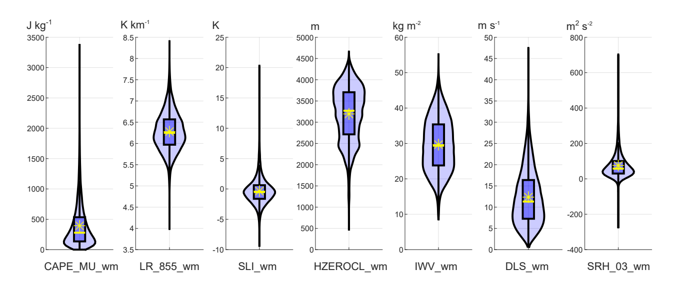
[3] Sofern nicht explizit anders angegeben handelt es sich in Kapitel 5.2.1 immer um das in Gleichung (4.12) beschriebene abstandsgewichtete Mittel der jeweiligen Variablen in der Zellumgebung (`wm`; erster Eintrag in der Spalte „Statistik“ in Tabelle E.1). Ab Kapitel 5.2.2 wird jeweils das statistische Maß der Umgebungsvariablen verwendet, das als zweiter Eintrag in der Spalte „Statistik“ in Tabelle E.1 geführt ist. Ausnahmen bilden die 850 hPa Temperatur und pseudopotentielle Temperatur, für die verschiedene statistische Maße verwendet werden, die jeweils explizit angegeben werden (`wm`, `max` oder `sdam`). Sofern nicht explizit anders angegeben wurden die Umgebungsvariablen zudem zeitlich über die komplette Lebensdauer der Zellobjekte gemittelt.

500 J kg^{-1} auf (93 % weniger als 1 000 J kg^{-1}; Abbildung 5.10a). Somit ist ein Großteil der Zellen eher im Bereich niedriger bis moderater CAPE aufgetreten. In Mitteleuropa sind hohe CAPE-Werte von weit über 1 000 J kg^{-1} im Vergleich zu den USA allgemein deutlich seltener (Brooks et al., 2003; Taszarek et al., 2020). Die Werte des SLI liegen meist um 0 K bzw. leicht im negativen (instabilen) Bereich, die der mitteltroposphärischen *Lapse Rate* $LR_{850-500hPa}$ befinden sich vornehmlich im bedingt labilen Bereich zwischen meist 5,5 und 7 K km^{-1} (vgl. Kapitel 2.1.2).

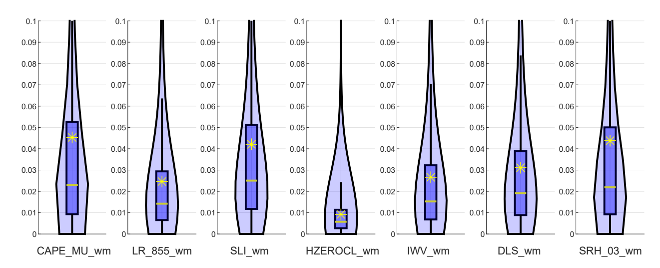
Bezüglich der vertikalen Windscherung weisen knapp 80 % der Zellobjekte eine niedrige bis moderate DLS von weniger als 18 m s^{-1} auf. Davon tritt nur jedes zwanzigste Zellobjekt bei einer $CAPE_{MU}$ von mehr als 1 000 J kg^{-1} bzw. bei einer maximalen Vertikalgeschwindigkeit $W_{MAX} = \sqrt{2\,CAPE_{MU}}$ von mehr als etwa 45 m s^{-1} auf (Abbildung 5.11a; vgl. Kapitel 2.1.2). Die Werte der SRH_{0-3km} liegen meist zwischen 0 und 200 m^2 s^{-2}. Solche Verteilungen der DLS und der SRH_{0-3km} sind in guter Übereinstimmung mit der im Vergleich zu Einzelzellen niedrigen beobachteten Häufigkeit von Superzellen in Mitteleuropa (vgl. Kapitel 2.2.3; Taszarek et al., 2020).

Gemäß der Clausius-Clapyeron-Gleichung (2.18), nach welcher der Sättigungsdampfdruck exponentiell mit zunehmender Temperatur steigt, finden sich hohe Werte des IWV bei hohen Werten der Temperatur T_{850hPa} (`wm`; Abbildung 5.11b). Der Anteil von Zellobjekten, die in warmer Luft mit $T_{850hPa} > 15$ °C registriert wurden, ist mit 11,5 % größer als der Anteil von 1,6 % der Zellen in kühler Frühlings- oder Herbstluft unter 0 °C. Insbesondere in warmer Luft finden sich Zellobjekte für einen weiten Wertebereich des IWV von etwa 20−45 kg m^{-2} (vgl. auch Abbildung 5.10a). Insgesamt treten konvektive Zellen für einen festen Wert von T_{850hPa} bevorzugt bei hohen Werten des IWV auf.

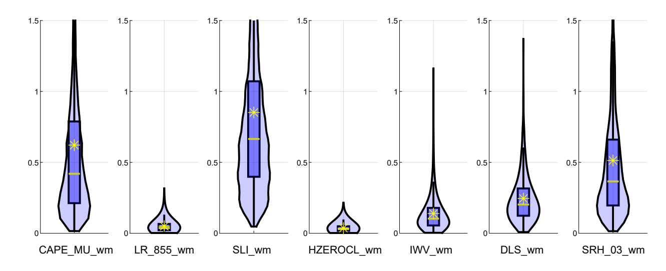
Die zeitliche Variabilität der Umgebungsvariablen während des Lebenszyklus der Zellobjekte im Datensatz ist in den meisten Fällen sehr gering (Abbildung 5.10b). Vergleicht man die jeweiligen Variabilitäten während eines



(a) Abstandsgewichtete Mittelwerte verschiedener Variablen in der Zellumgebung, zusätzlich zeitlich über den jeweiligen Lebenszyklus der Zellobjekte gemittelt (38 553 Objekte).

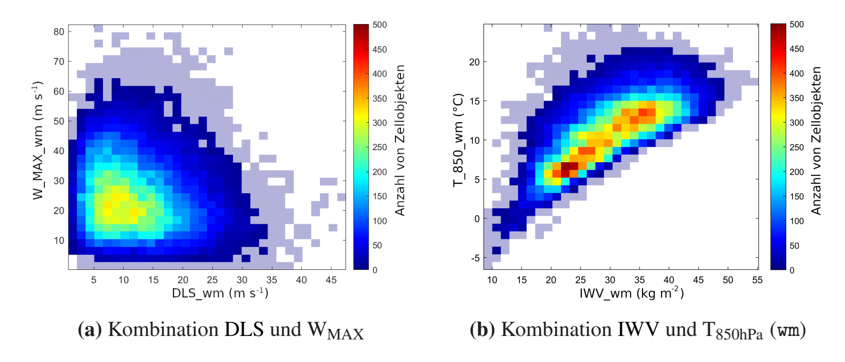


(b) Wie (a), nur wird statt des zeitlichen Mittels die entsprechende zeitliche Standardabweichung betrachtet (38 553 Objekte).



(c) Wie (b), nur mit einer Beschränkung der Stichprobe auf Zellobjekte mit einer Lebensdauer von mehr als 60 min (1 096 Objekte).

Abbildung 5.10: Häufigkeitsverteilungen einer Auswahl atmosphärischer Variablen: $CAPE_{MU}$ ($J\,kg^{-1}$), $LR_{850-500hPa}$ ($K\,km^{-1}$), SLI (K), 0 °C-Grenze (m), IWV ($kg\,m^{-2}$), DLS ($m\,s^{-1}$), SRH_{0-3km} ($m^2\,s^{-2}$). Der Interquartilsbereich ist durch Boxen, der Median durch die gelbe Linie und das arithmetische Mittel durch einen gelben Stern hervorgehoben. Die Kerndichteschätzung der Verteilung nach Parzen (1962) verwendet einen Gaußkern.

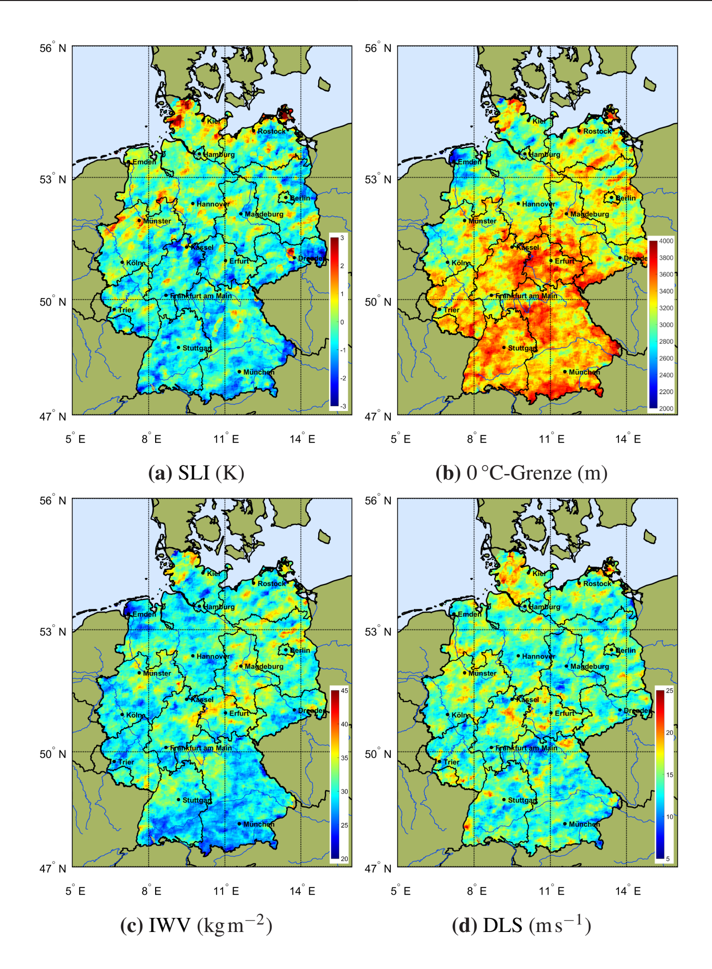


(a) Kombination DLS und W_{MAX} **(b)** Kombination IWV und T_{850hPa} (wm)

Abbildung 5.11: Kombinierte Häufigkeitsverteilungen für die Umgebungsbedingungen aller 38 553 Zellobjekte (a) für die Kombination DLS und W_{MAX} und (b) die Kombination IWV und T_{850hPa} (wm). Wertebereiche mit weniger als zehn zugeordneten Zellobjekten sind transparent dargestellt.

Lebenszyklus (ausgedrückt durch die Standardabweichung der Werte von allen jeweiligen Detektionszeitpunkten der Zellobjekte) mit der Variabilität der Mittelwerte im Datensatz (also der Standardabweichung der Werte aus Abbildung 5.10a), so liegt das Verhältnis aus diesen beiden Größen meist unter 0,1. Allerdings stellt der Großteil der Zellobjekte kurzlebige konvektive Zellen dar. Je höher die Lebensdauer ist, desto größer ist die Variabilität während des Lebenszyklus für viele der Umgebungsvariablen (Abbildung 5.10c). Dabei liegt für einige Variablen wie die $CAPE_{MU}$, den SLI und die SRH_{0-3km} das Verhältnis aus der Variabilität während des Lebenszyklus und der Variabilität der Mittelwerte im Datensatz für einige Zellobjekte bei mehr als 0,5. Hier treten also recht große Variabilitäten der Umgebungsvariablen über einen Lebenszyklus der Zellobjekte auf (im Vergleich zur generellen Variabilität der Mittelwerte im Datensatz, s. o.), die mit der Veränderung der Umgebungsbedingungen durch die Modelldynamik und -physik bzw. durch die Assimilation zusammenhängen (vgl. Kapitel 4.2). Dieselben Schlüsse lassen sich ziehen, wenn man für das Mittel in der Zellumgebung statt dem abstandsgewichteten das arithmetische Mittel oder den Median verwendet. Das Ergebnis

zur höheren Variabilität einiger Umgebungsvariablen während des Lebenszyklus unterscheidet sich dabei von anderen Studien (vgl. Sun et al., 2014; Zöbisch et al., 2020). Die höhere Variabilität ist aufgrund der Größe des Umgebungsradius von meist 27 und 32 km bei der Zusammenführung der Daten (vgl. Kapitel 4.3.4) möglicherweise darauf zurückzuführen, dass die betroffenen Variablen auf einer kleinen räumlichen Skala in der Größenordnung des Umgebungsradius variieren. Bei einigen thermodynamischen Größen wie z. B. dem LI oder der CAPE kommt zusätzlich die Sensitivität ihrer Werte bezüglich der Temperatur- und Feuchtewerte in den untersten Troposphärenschichten hinzu. Diese wirken sich auf die Höhe des HKN und damit auf die Stabilität der Luftschichtung aus (Lee, 2002; Miller und Mote, 2018). Die folgenden Analysen verwenden dennoch zur Reduzierung der Dimension nur einen für den Lebenszyklus jedes Zellobjekts repräsentativen Wert für die Umgebungsvariablen (z. B. den Mittelwert des Lebenszyklus oder den Wert zum Zeitpunkt der ersten Detektion).
Die Umgebungsbedingungen, die während des Auftretens konvektiver Zellen vorherrschen, können auch räumlich variabel sein. Da in der vorliegenden Arbeit Zellen von April bis September in die Analysen eingehen, schwankt beispielsweise die 0 °C-Grenze zwischen 2 000 und mehr als 4 000 m, wobei die Zellobjekte im Südosten des Lands prinzipiell höhere Werten aufweisen als im Nordwesten (Abbildung 5.12b). Dies lässt sich hauptsächlich auf die generelle mittlere Luftmassenverteilung im Sommerhalbjahr über Deutschland zurückführen, die z. B. durch einen Gradienten der 850 hPa Temperatur von Nord(west) nach Süd(ost) charakterisiert ist (Abbildung D.3). Viele Umgebungsvariablen wie z. B. der SLI, der IWV oder die DLS zeigen keinen großskaligen horizontalen Gradienten (Abbildungen 5.12a,c,d). Die Werte des SLI liegen, bis auf den Nordwesten des Lands, meist um 0 K oder leicht im negativen Bereich. Die Werte der DLS lassen gar keinen großskaligen Gradienten erkennen und variieren zwischen moderaten Werten von $10-20\,\mathrm{m\,s^{-1}}$. Die Werte des IWV sind insbesondere im äußersten Süden im Bereich des



(a) SLI (K) **(b)** 0 °C-Grenze (m)

(c) IWV ($\mathrm{kg\,m^{-2}}$) **(d)** DLS ($\mathrm{m\,s^{-1}}$)

Abbildung 5.12: Räumliche Verteilung der mit den Zellobjekten assoziierten Umgebungsbedingungen im Untersuchungszeitraum der Sommerhalbjahre 2011 – 2016, beispielhaft für (a) den SLI, (b) die 0 °C-Grenze, (c) den IWV und (d) die DLS. Ähnlich zu Abbildung 5.1a wird für jeden Gitterpunkt eines $1 \times 1\,\mathrm{km}^2$-Gitters der Mittelwert einer Umgebungsvariablen über all diejenigen Zellobjekte bestimmt, deren Polygone den Gitterpunkt einschließen. Jedes Polygon enthält dabei einen festen Wert für jede Umgebungsvariable, und zwar den abstandsgewichteten Mittelwert der Variablen in der Zellumgebung, zusätzlich zeitlich über den jeweiligen Lebenszyklus der Zellobjekte gemittelt (vgl. Abbildung 5.10a). Um der geringen Objektanzahl Rechnung zu tragen, erfolgt abschließend eine Glättung des so bestimmten Felds über 7×7 Gitterpunkte.

höher gelegenen Alpenvorlands, des Schwarzwalds, der Baar und der Schwäbischen Alb mit $20-30\,\mathrm{kg\,m^{-2}}$ tendenziell etwas niedriger als im Rest des Lands. Da die Anzahl von Zellobjekten pro Gitterpunkt meist zwischen 5 und 25 liegt (vgl. Abbildung 5.1a), dürfen regionale Unterschiede in den Werten der Umgebungsvariablen auf räumlichen Skalen der Größenordnung von $\mathcal{O}(100\,\mathrm{km})$ nicht überinterpretiert werden.

5.2.2 Korrelationsanalyse und Clustering der Umgebungsvariablen

Die Kombination der Lebenszyklen der Zellobjekte mit den Umgebungsvariablen ermöglicht eine objektbasierte Untersuchung der Korrelationen jeweils zweier unterschiedlicher atmosphärischer Variablen. Basierend auf dem Wissen über die meteorologischen und statistischen Zusammenhänge aus den folgenden Abschnitten können in Kapitel 6 bessere Entscheidungen für durchzuführende Modellstudien getroffen werden. Als Übersicht über verschiedene paarweise Korrelationen dient die symmetrische und positiv semidefinite Korrelationsmatrix (vgl. Kapitel 3.1), wobei in Abbildung 5.13 die Rangkorrelationen nach Spearman (r_S) dargestellt sind. Werte des (linearen) Korrelationskoeffizienten nach Pearson r_P sind meist nur geringfügig kleiner (nicht gezeigt). Deutliche Unterschiede zwischen den beiden Korrelationskoeffizienten treten bei Korrelationen mit Variablen auf, deren Verteilungen eine große Schiefe haben, z. B. bei Korrelationen mit der $\mathrm{BRN_{MU}}$, der $\mathrm{CAPE_{MU}}$ oder dem SCP. Aufgrund der großen Stichprobe ($N = 38\,553$) sind die meisten Korrelationen statistisch signifikant. Über die (lineare) Hauptkomponentenanalyse (Kapitel 3.1.2) können statistische Zusammenhänge zweier Variablen veranschaulicht werden.

Wie erwartet, zeigt der Betrag des Spearman'schen Korrelationskoeffizienten r_S jeweils zwischen zwei dynamischen oder zwei thermodynamischen Größen häufig hohe Werte. Die erste Komponente der Hauptkomponentenanalyse zwischen der $\mathrm{SRH_{0-3km}}$ und der DLS erklärt bereits mehr als 77 %

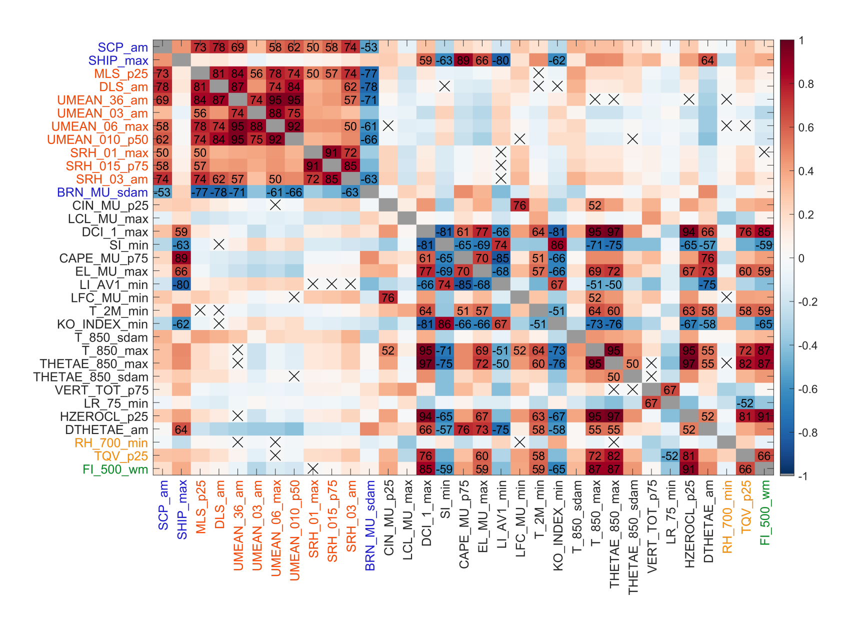


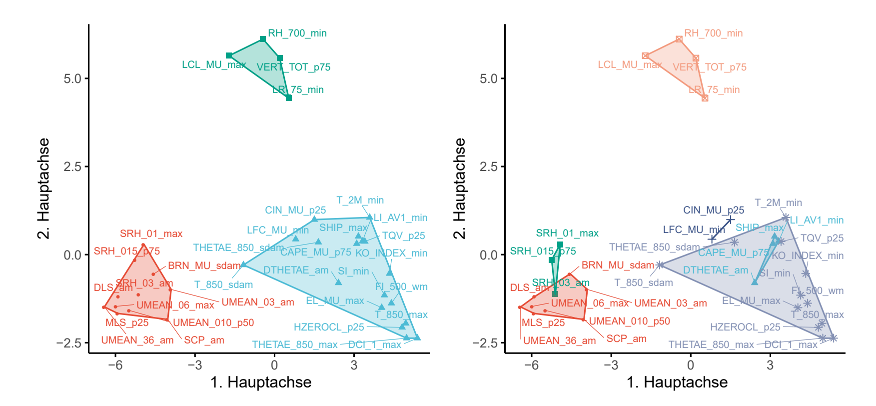
Abbildung 5.13: Korrelationsmatrix des Spearman'schen Rang-Korrelationskoeffizienten r_S für eine Auswahl von Umgebungsvariablen. Zur Erläuterung der verwendeten Abkürzungen sei auf Tabelle E.1 verwiesen. Hohe (Anti-)Korrelationen sind in kräftigen Farbtönen dargestellt (rot: $r_S > 0$, blau: $r_S < 0$), niedrige in blassen. Korrelationen über 0,5 sind zusätzlich als Zahlenwerte eingetragen (%). Statistisch insignifikante Korrelationen sind durch ein Kreuz markiert (Signifikanzniveau $p = 0{,}01$). Die Akronyme der Umgebungsvariablen sind wie folgt eingefärbt: dynamische Größen (rot), thermodynamische Größen und Grenzhöhen (schwarz), reine Feuchtegrößen (ocker), kombinierte Kenngrößen (blau), 500 hPa Geopotential (grün).

der Gesamtstreuung bei $r_P = 0{,}52$ und $r_S = 0{,}62$ (Abbildung D.4a). Noch stärker korreliert die DLS jedoch mit dem mittleren mitteltroposphärischen Horizontalwind, beispielsweise mit dem mittleren Wind zwischen 3 und 6 km Höhe $\overline{\mathrm{U}}_{3-6\,\mathrm{km}}$ ($r_S = 0{,}87$). Diese hohe Korrelation spiegelt wider, dass der Wert der DLS maßgeblich durch den Betrag des Winds in 6 km Höhe bestimmt ist. Die paarweisen Korrelationen zwischen den verschiedenen Varianten der SRH sind ebenfalls hoch, auch zwischen $\mathrm{SRH}_{0-1\mathrm{km}}$ und $\mathrm{SRH}_{0-3\mathrm{km}}$. Eine

mögliche Interpretation ist, dass bereits im untersten Kilometer über Grund häufig eine merkliche Richtungsscherung vorherrscht, welche entscheidend für die Verfügbarkeit von *Streamwise Vorticity* ist (vgl. Kapitel 2.2.3).
Die erste Komponente der Hauptkomponentenanalyse zwischen W_{MAX} und dem LI_{100hPa} erklärt mehr als 90 % der Gesamtstreuung bei $r_P = -0{,}81$ (Abbildung D.4b). Zudem korreliert der LI_{100hPa} stark mit weiteren thermodynamischen Größen, welche die thermische Instabilität beschreiben, wie dem SI ($r_S = 0{,}74$), dem KO-Index ($r_S = 0{,}67$) oder $\Delta\theta_{ps}$ ($r_S = 0{,}76$). Die Stärke des statistischen Zusammenhangs unterscheidet sich dabei teilweise von den Ergebnissen basierend auf Radiosondendaten von Mohr und Kunz (2013). Dies könnte damit zusammenhängen, dass dort zur Berechnung konvektiver Indizes die Radiosondenmessungen um 12 UTC verwendet wurden, während in die vorliegenden Korrelationsanalyse Werte zum jeweiligen Detektionszeitpunkt im unmittelbaren Umfeld einer konvektiven Zelle eingehen. Allgemein treten hohe Korrelationen zwischen zwei Stabilitätsmaßen unabhängig von der Art der Instabilität auf, die sie beschreiben (bedingte, latente, potentielle Instabilität; vgl. Kapitel 2.1.2).
Der kombinierte Index SCP, der sich allgemein multiplikativ aus der $CAPE_{MU}$, der DLS und der SRH_{0-3km} zusammensetzt, korreliert mehr mit dynamischen Größen, während der SHIP, in den als einzige dynamische Größe die DLS eingeht (vgl. Kapitel 2.3), verstärkt mit thermodynamischen Größen korreliert. Die BRN_{MU} (vgl. Kapitel 2.2.2) wiederum ist mehr von der mitteltroposphärischen Dynamik als der Thermodynamik geprägt, wie die hohen Korrelationen mit der vertikalen Windscherung und den mittleren Winden andeuten.
Eine niedrige Korrelation zeigt sich beispielsweise zwischen LI_{100hPa} und DLS ($r_S = 0{,}15$, $r_P = 0{,}19$), welche den meteorologisch plausiblen geringen Zusammenhang zwischen vertikaler Windscherung und thermischer Instabilität widerspiegelt. Die 3σ-Ellipse (vgl. Kapitel 3.1.2) in der

Hauptkomponentenanalyse dieser beiden Umgebungsvariablen ist daher der Form eines Kreises recht nahe (Abbildung D.4c).

Um ein objektives, detailliertes Bild über miteinander korrelierende Variablen zu erhalten, findet ein *k-Medoids*-Clustering Anwendung (Kapitel 3.2). Als Distanzmaß dient $d_C = 1 - |r_S|$, d. h. stark (anti-)korrelierende Variablen haben in dieser Metrik einen geringen Abstand zueinander (niedrige Dissimilation; z. B. Van der Laan et al., 2003). Für eine Anzahl von $N_C = 3$ Clustern (Abbildung 5.14a) findet der PAM-Algorithmus einen dynamischen Cluster mit $\overline{\mathbf{U}}_{3-6\mathrm{km}}$ als Medoid bei einem mittleren Silhouettenkoeffizienten von $\bar{s}_1 = 0{,}55$ (mittlere Strukturierung). Dieser Cluster befindet sich bei negativen Werten der ersten Hauptachse des mittels multidimensionaler Skalierung geschätzten Raums (vgl. Kapitel 3.2). Der zweite und dritte Cluster setzen sich aus thermodynamischen und Feuchtegrößen zusammen, mit dem $\mathrm{DCI}_{100\mathrm{hPa}}$ und dem *Vertical Totals* (VT) als Medoiden und nur geringer Strukturierung ($\bar{s}_2 = 0{,}38$; $\bar{s}_3 = 0{,}24$). Der eine befindet sich bei positiven Werten der ersten Hauptachse, während sich der andere von ihm und dem dynamischen Cluster im Beitrag der zweiten Hauptachse maßgeblich unterscheidet. Alle 33 einzelnen Silhouetten $s_k^{(q)}$ sind positiv, sodass von einem erfolgreichen Clustering gesprochen werden kann.
Für $N_C = 6$ Cluster erfolgt eine Aufteilung des Clusters, der sich aus dynamischen Größen zusammensetzt, in die drei SRH-Variablen und die übrigen dynamischen Umgebungsvariablen (Abbildung 5.14b). Zudem erfolgt eine Separation zweier weiterer kleiner thermodynamischer Cluster: Der eine besteht hauptsächlich aus Variablen, welche die atmosphärische Stabilität beschreiben, den anderen bilden die $\mathrm{CIN}_{\mathrm{MU}}$ und das $\mathrm{NFK}_{\mathrm{MU}}$, welche auch physikalisch direkt miteinander assoziiert sind (vgl. Kapitel 2.1.2). Die mittleren Silhouettenkoeffizienten variieren zwischen $\bar{s}_5 = 0{,}16$ und $\bar{s}_3 = 0{,}65$ (sehr geringe bis mittlere Strukturierung). Die Silhouetten von vier Variablen im großen thermodynamischen Cluster sind allerdings leicht negativ und könnten demnach auch dem thermodynamischen Cluster um die $\mathrm{CAPE}_{\mathrm{MU}}$ zugeordnet



(a) Drei Cluster mit den Medoiden $\overline{\mathbf{U}}_{3-6\mathrm{km}}$, $\mathrm{DCI}_{100\mathrm{hPa}}$ und VT

(b) Sechs Cluster mit den Medoiden $\overline{\mathbf{U}}_{3-6\mathrm{km}}$, $\mathrm{DCI}_{100\mathrm{hPa}}$ und VT, $\mathrm{CAPE}_{\mathrm{MU}}$, $\mathrm{SRH}_{0-1{,}5\mathrm{km}}$ und $\mathrm{NFK}_{\mathrm{MU}}$

Abbildung 5.14: Darstellung der (a) drei und (b) sechs verschiedenen Cluster von Umgebungsvariablen, die ein *k-Medoids*-Clusterverfahren identifiziert. Als Distanzmaß dient $d_C = 1 - |r_S|$. Dargestellt ist die Projektion der Cluster auf die ersten beiden Hauptachsen des hochdimensionalen Eigenschaftsraums. Die erste Hauptachse erklärt 52,1 %, die zweite 15,7 % der beobachteten Variabilität.

werden. Die Auftrennung ist hier folglich nicht so eindeutig wie die des dynamischen Clusters. Der Silhouettenkoeffizient über den gesamten Datensatz $\overline{S}$ ist für $N_C = 3$ mit $\overline{S} = 0{,}42$ etwas höher als für $N_C = 6$ mit $\overline{S} = 0{,}37$. Für $N_C \in [2;10]$ erreicht $N_C = 3$ den höchsten Wert und stellt damit die beste Strukturierung des Datensatzes dar.

Führt man ein *k-Medoids*-Clustering mit weiteren Umgebungsvariablen durch, findet man beispielsweise, dass auch andere *Lapse Rates*, die relative Luftfeuchtigkeit auf anderen Niveaus und der *Total Totals* (TT) einem Cluster um den VT zugeordnet würden. Eine weitere Erhöhung der Anzahl von Clustern N_C separiert im dynamischen Cluster nieder- und mitteltroposphärische Größen. Das diskutierte Clustering bedeutet nicht, eine optimale Zuordnung getroffen zu haben. Vielmehr gibt die Korrelations- und Clusteranalyse einen

anschaulichen und objektiven Überblick über den Zusammenhang verschiedener Umgebungsvariablen. Darüber hinaus ist sie für die Auswahl der Prädiktoren bei den Untersuchungen von Vorhersageverfahren für die Lebensdauer oder die maximale Fläche konvektiver Zellen nützlich (s. Kapitel 6.1.1, 6.3 und 6.4).

5.3 Einfluss von Umgebungsbedingungen auf Zellattribute

5.3.1 Univariate Analysen

Unterscheidungsvermögen der Umgebungsvariablen

Das Unterscheidungsvermögen (*Discrimination*) der Umgebungsvariablen hinsichtlich unterschiedlicher Werte für die Lebensdauer oder maximale Fläche der Zellobjekte (Zellattribute) wird mit Hilfe von Häufigkeitsverteilungen untersucht. Mittels der Kerndichteschätzung nach Parzen (1962) unter Verwendung eines Gaußkerns erfolgt eine Schätzung der Verteilungsfunktionen der Umgebungsvariablen für unterschiedliche Gruppen von Zellobjekten, welche anschließend verglichen werden. Zum Vergleich zweier Verteilungsfunktionen ist deren Überlappung bzw. *Overlap* (OLP) als Anteil der sich überlagernden Flächen unter den Graphen nützlich, oder der Unterscheidungsfaktor bzw. *Discrimination Factor* (DIS), welcher durch $DIS = 1 - OLP$ gegeben ist. Zusätzlich werden weitere Unterscheidungsmaße zwischen zwei Verteilungsfunktionen verwendet, beispielsweise die maximale *True Skill Statistic* (TSS) oder der maximale *Critical Success Index* (CSI; vgl. Tabelle 3.2; z. B. Czernecki et al., 2019). Dabei wird iterativ für verschiedene Variablentrennwerte (*Cutting Points*; z. B. 100 verschiedene im Wertebereich) der jeweilige Score bestimmt, womit durch einen Wertevergleich der maximale Score für einen optimalen Variablentrennwert identifiziert wird. Eine Untersuchung der Sensitivität der Scores bezüglich der Anzahl von Variablentrennwerten folgt weiter unten in diesem Kapitel.

Aufgrund der Schiefe des Datensatzes bezüglich der Zellattribute ist es wichtig zu berücksichtigen, dass dieses Ungleichgewicht den *CSI* beeinflusst, die *TSS* per definitionem hingegen nicht, weswegen beide Gütemaße betrachtet werden. Die Trefferrate (H) und die Fehlalarmrate (F) geben in diesem Fall den Anteil der Zellobjekte der jeweiligen Gruppe an, der sich rechts (links) des optimalen Trennwerts einer zum Zellattribut proportionalen (antiproportionalen) Umgebungsvariablen befindet. Die *TSS* entspricht wie üblich deren Differenz von Treffer- und Fehlalarmrate. Der *CSI* bestimmt sich als Anteil der korrekten Zuordnungen der Gruppe mit hohen Werten für die Zellattribute an allen Zuordnungen bis auf die korrekten Zuordnungen der Gruppe mit niedrigen Werten für die Zellattribute.
Die folgenden Analysen zum Unterscheidungsvermögen beschränken sich auf zwei Gruppen an Zellobjekten, beispielsweise die Fähigkeit einer Umgebungsvariablen zwischen Zellen mit kurzer und langer Lebensdauer zu unterscheiden[4]. Die Resultate sind inhärent abhängig von der Wahl der Klassentrennwerte (vgl. Kapitel 3.6.1) zwischen den zwei Gruppen für die Lebensdauer (Klassentrennwert τ) bzw. die maximale Zellfläche (Klassentrennwert χ). Deswegen folgt im Anschluss eine kurze Diskussion der diesbezüglichen Sensitivität.

Unterscheidungsvermögen in Bezug auf die Lebensdauer

Für alle 83 berechneten Umgebungsvariablen werden mit allen neun unterschiedlichen statistischen Maßen aus der Statistik der Zellumgebung (vgl. Kapitel 4.3.4) für zwei Lebensdauer-Gruppen (kurz/lang) mit verschiedenen Klassentrennwerten der Lebensdauer τ mehrere

[4] Aufgrund der Schiefe der Verteilung der Lebensdauer im Datensatz unterscheiden sich bei drei oder mehr Gruppen die Verteilungen der Umgebungsvariablen in der Gruppe der Zellobjekte mit der kürzesten Lebensdauer kaum von derjenigen, die man bei lediglich zwei Gruppen erhält.

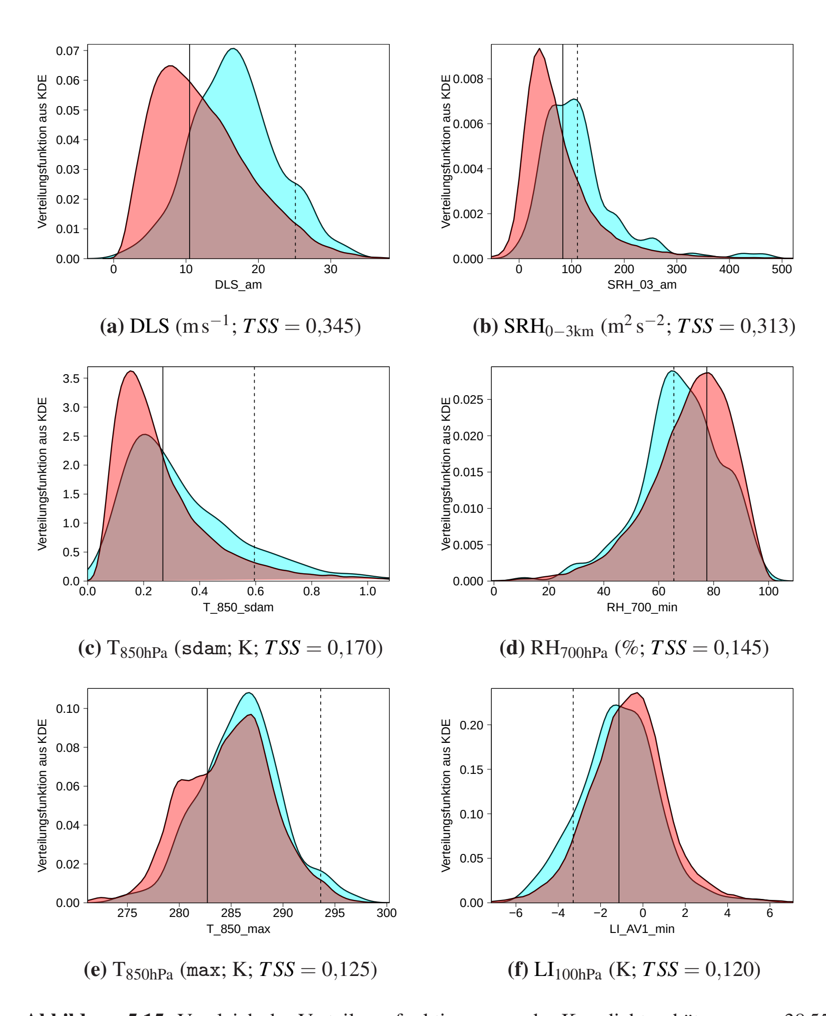
Scores berechnet. Zur Reduzierung der sehr großen Anzahl von Variablen ($9 \cdot 83 = 747$) erfolgt eine absteigende Sortierung nach den Werten für die TSS. Alle Größen, die eine $TSS < 0{,}1$ für $\tau = 100\,\text{min}$ aufweisen, werden aussortiert. Dieses Entscheidungskriterium resultiert aus vielen Untersuchungen mit verschiedenen Klassentrennwerten und dem Ziel, eine adäquate Anzahl von Umgebungsvariablen für die weiteren Analysen zu verwenden. Erreichen mehrere statistische Maße einer Variablen (z. B. der arithmetische Mittelwert und das Minimum der Variablen in der Zellumgebung) $TSS \geq 0{,}1$, so wird lediglich diejenige Größe beibehalten, welche die höhere TSS aufweist (vgl. Kapitel 4.3.4). Für fast alle Variablen korrelieren die statistischen Maße aus der Umgebungsstatistik sehr stark (nicht gezeigt) und decken somit den gleichen Gehalt an Information ab. Basierend auf der Anwendung dieser Auswahlkriterien ergeben sich 28 verbleibende Variablen. Diese werden um fünf weitere Variablen ergänzt: HKN_{MU}, $\Delta\theta_{\text{ps}}$, IWV, $\text{T}_{850\text{hPa}}$ (`sdam`) und $\theta_{\text{ps,850hPa}}$ (`sdam`). Die Hinzunahme der ersten drei Variablen ist in physikalischen Überlegungen begründet, um die Diversität der Auswahl zu erhöhen, obwohl deren $TSS < 0{,}1$ ist. Die letzten beiden weisen eine $TSS > 0{,}1$ auf und korrelieren kaum mit der ausgewählten $\text{T}_{850\text{hPa}}$ (`max`) bzw. $\theta_{\text{ps,850hPa}}$ (`max`) oder anderen Größen der Auswahl (vgl. Abbildung 5.13). Aus meteorologischer Sicht können beide Größen ein Indikator für scharfe Luftmassengrenzen sein.

Die höchsten TSS-Werte erreichen fast ausschließlich dynamische Variablen (Tabelle 5.1). Die CSI-Werte sind bei der Wahl eines hohen Klassentrennwerts von $\tau = 100\,\text{min}$ generell sehr niedrig, da es absolut gesehen viel mehr falsche Zuordnungen von Zellobjekten mit kurzer Lebensdauer gibt (38 335 Objekte mit einer kurzen Lebensdauer von weniger als 100 min verglichen mit 218 mit einer langen Lebensdauer von mehr als 100 min). Im Hinblick auf das Unterscheidungsvermögen erreicht der SCP die höchste TSS, die DLS liegt auf Rang 3 und die $\text{SRH}_{0-3\text{km}}$ auf Rang 6. Exemplarisch sind die DLS-Werte von etwa 89 % aller Zellobjekte mit einer langen

Lebensdauer höher als der Variablentrennwert DLS $> 10{,}5\,\mathrm{m\,s^{-1}}$, der zu obiger TSS führt (Abbildung 5.15a). Damit liegt die Trefferrate für ein langlebiges Zellobjekt bei DLS-Werten, die größer als dieser Variablentrennwert sind, bei $H = 0{,}89$. Gleichzeitig liegen die DLS-Werte von etwa 55 % aller Zellobjekte mit einer kurzen Lebensdauer höher als der Variablentrennwert DLS $> 10{,}5\,\mathrm{m\,s^{-1}}$, was $F = 0{,}55$ und damit gerundet $TSS = 0{,}35$ bedeutet. Des Weiteren rangieren $\mathrm{T_{850hPa}}$ (sdam), $\mathrm{RH_{700hPa}}$ und der $\mathrm{LI_{100hPa}}$ beispielsweise auf den Plätzen 10, 14 und 18. Die Werte der TSS sind insgesamt denen des Unterscheidungsfaktors DIS sehr ähnlich. Bei Variablen mit einer scharfen Begrenzung des Wertebereichs und einer starken Schiefe hin zu dieser Begrenzung (z. B. 0 bei dem SCP, der BRN und dem SHIP) führt die implementierte Berechnungsmethodik für den OLP und die DIS auf Basis der Kerndichteschätzung zu wenig sinnvollen Werten, weswegen diese Gütemaße für solche Umgebungsvariablen nicht betrachtet werden (Tabelle 5.1).
Allerdings erreicht keine der Umgebungsvariablen eine trennscharfe Unterscheidung (Abbildung 5.15). Die Dominanz der dynamischen Variablen lässt darauf schließen, dass diese aufgrund ihres Einflusses auf die Organisationsform von konvektiven Zellen (vgl. Kapitel 2.2) das beste Unterscheidungsvermögen bezüglich der Lebensdauer besitzen. Auch die mitteltroposphärische Feuchte $\mathrm{RH_{700hPa}}$ sowie einige thermodynamische Variablen wie z. B. der KO-Index, die statistisch nur schwach mit dynamischen Variablen zusammenhängen (vgl. Kapitel 5.2.2), weisen maximal als mäßig gut einzuordnende Werte der Gütemaße auf (im Vergleich zu den Werten der dynamischen Variablen). Allerdings liegt die Vermutung nahe, dass eine Kombination von mehreren Umgebungsvariablen in der Lage sein könnte noch bessere Unterscheidungen vorzunehmen (s. Kapitel 5.3.2 und 6).
Wendet man dieselbe Technik statt auf Umgebungsvariablen auf die Zellfläche zum Zeitpunkt der dritten oder vierten Detektion an, so erhält man $TSS = 0{,}31$ bzw. $TSS = 0{,}30$ mit den optimalen Variablentrennwerten $A_Z(t = 12\,\mathrm{min}) = 25{,}5\,\mathrm{km^2}$ und $A_Z(t = 17\,\mathrm{min}) = 29{,}3\,\mathrm{km^2}$ (nicht gezeigt). Die besten sechs Umgebungsvariablen haben somit ein besseres

Tabelle 5.1: Übersicht über verschiedene Maße des Unterscheidungsvermögens der 15 Umgebungsvariablen mit der höchsten TSS für den Klassentrennwert der Lebensdauer von $\tau = 100\,\text{min}$. Zusätzlich ist der optimale Variablentrennwert, der zur gelisteten TSS führt, angegeben.

Umgebungsvariable	*TSS*	*H*	*F*	*CSI*	*OLP*	*DIS*	Variablentrennwert	Einheit
SCP	0,37	0,85	0,48	0,01	—	—	0,05	—
MLS	0,35	0,74	0,39	0,02	0,68	0,32	8,9	$\mathrm{m\,s^{-1}}$
DLS	0,35	0,89	0,55	0,01	0,67	0,33	10,5	$\mathrm{m\,s^{-1}}$
$\overline{\mathbf{U}}_{3-6\mathrm{km}}$	0,32	0,80	0,48	0,01	0,70	0,30	11,6	$\mathrm{m\,s^{-1}}$
$\overline{\mathbf{U}}_{0-10\mathrm{km}}$	0,32	0,76	0,44	0,01	0,71	0,29	12,4	$\mathrm{m\,s^{-1}}$
$\mathrm{SRH}_{0-3\mathrm{km}}$	0,31	0,64	0,33	0,01	0,70	0,30	83,1	$\mathrm{m^2\,s^{-2}}$
$\mathrm{BRN}_{\mathrm{MU}}$	0,27	0,77	0,50	0,01	—	—	8,75	—
$\overline{\mathbf{U}}_{0-6\mathrm{km}}$	0,26	0,75	0,49	0,01	0,77	0,23	10,2	$\mathrm{m\,s^{-1}}$
$\mathrm{SRH}_{0-1,5\mathrm{km}}$	0,20	0,67	0,47	0,01	0,81	0,20	50,4	$\mathrm{m^2\,s^{-2}}$
$\mathrm{T}_{850\mathrm{hPa}}$ (`sdam`)	0,17	0,54	0,37	0,01	—	—	0,27	K
SHIP	0,16	0,57	0,41	0,01	—	—	0,09	—
$\mathrm{SRH}_{0-1\mathrm{km}}$	0,16	0,61	0,44	0,01	0,84	0,16	62,7	$\mathrm{m^2\,s^{-2}}$
KO-Index	0,16	0,76	0,60	0,01	0,85	0,15	−3,4	K
$\mathrm{RH}_{700\mathrm{hPa}}$	0,15	0,76	0,61	0,01	0,86	0,14	77,5	%
$\mathrm{DCI}_{100\mathrm{hPa}}$	0,13	0,64	0,51	0,01	0,90	0,11	21,0	K



(a) DLS ($\mathrm{m\,s^{-1}}$; $TSS = 0{,}345$)

(b) $\mathrm{SRH_{0-3km}}$ ($\mathrm{m^2\,s^{-2}}$; $TSS = 0{,}313$)

(c) $\mathrm{T_{850hPa}}$ (`sdam`; K; $TSS = 0{,}170$)

(d) $\mathrm{RH_{700hPa}}$ (%; $TSS = 0{,}145$)

(e) $\mathrm{T_{850hPa}}$ (`max`; K; $TSS = 0{,}125$)

(f) $\mathrm{LI_{100hPa}}$ (K; $TSS = 0{,}120$)

Abbildung 5.15: Vergleich der Verteilungsfunktionen aus der Kerndichteschätzung von 38 553 Zellobjekten mit kurzer (rot; 38 335) und langer (blau; 218) Lebensdauer (Klassentrennwert der Lebensdauer: $\tau = 100\,\mathrm{min}$) für sechs verschiedene Umgebungsvariablen. Die vertikale durchgezogene Linie illustriert den optimalen Variablentrennwert basierend auf der TSS, die gestrichelte denjenigen, der den höchsten *CSI*-Wert hervorbringt. Die Verteilungsfunktion aus der Kerndichteschätzung ist so normiert, dass die Fläche unter dem Graphen 1 ergibt, sodass ihre Funktionswerte einheitenabhängig sind.

Unterscheidungsvermögen als die Zellfläche zum Zeitpunkt 10 oder 15 min nach der ersten Detektion, zumindest für einen hohen Klassentrennwert von $\tau = 100$ min. Für niedrigere Werte von τ unter etwa $80-85$ min haben die dynamischen Variablen ein niedrigeres Unterscheidungsvermögen als die Zellfläche zum Zeitpunkt 10 oder 15 min nach der ersten Detektion. Dies deutet insgesamt darauf hin, dass man rein anhand der Umgebungsbedingungen die Lebensdauer einer konvektiven Zelle ähnlich gut abschätzen kann wie mit Hilfe des Wissens über das anfängliche Wachstum der Zelle. Die Berücksichtigung einiger Umgebungsvariablen in einem Vorhersageverfahren für das *Nowcasting* – zusätzlich zu dem Wissen über den Verlauf der Zellattribute – könnte daher gewinnbringend sein, wie in späteren Modellstudien untersucht wird (Kapitel 6.3).

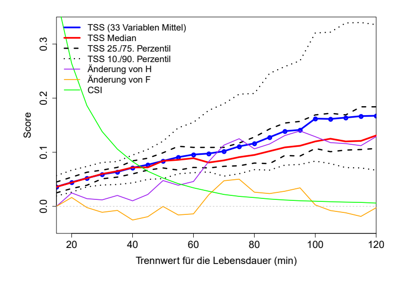
Unterscheidungsvermögen in Bezug auf die maximale Zellfläche

Im Gegensatz zur Lebensdauer erreichen bei der Unterscheidung von kleinen und großen maximalen Zellflächen (beispielhafter Klassentrennwert: $\chi = 80\,\mathrm{km}^2$; 38 264 kleine, 289 große Zellobjekte) thermodynamische Parameter die höchsten TSS-Werte (Tabelle 5.2). Qualitativ ändert sich die Reihenfolge der Umgebungsvariablen bei der Verwendung anderer Klassentrennwerte kaum. Besonders die konvektiven Indizes zur Beschreibung der Instabilität zeigen das beste Unterscheidungsvermögen, wie der $\mathrm{DCI}_{100\mathrm{hPa}}$ auf Platz 2, mit ihm in Verbindung stehend der $\mathrm{LI}_{100\mathrm{hPa}}$ auf Platz 4 oder der KO-Index auf Platz 6. Höhere Instabilitätswerte deuten auf ein höheres Maß an freier Konvektion hin, welches ein schnelleres Wachstum konvektiver Zellen begünstigt. Aber auch die vertikale Windscherung in Form der DLS weist ein gewisses Unterscheidungsvermögen auf. Insgesamt sind die TSS-Werte bei einer ähnlichen Anzahl von Zellobjekten in den beiden Gruppen etwas geringer als bei der Unterscheidung der Lebensdauer.

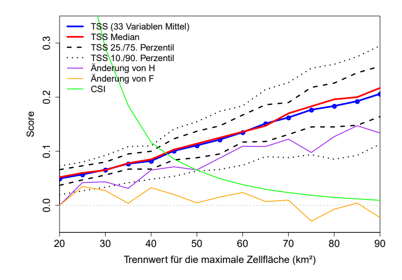
Die Zellflächen 10 und 15 min nach der ersten Detektion weisen sehr hohe Werte von $TSS = 0{,}61$ bzw. $TSS = 0{,}67$ mit den optimalen Variablentrennwerten $A_Z(t = 12\,\mathrm{min}) = 34{,}6\,\mathrm{km}^2$ und $A_Z(t = 17\,\mathrm{min}) = 41{,}0\,\mathrm{km}^2$ auf (nicht

Tabelle 5.2: Wie Tabelle 5.1, nur für den Klassentrennwert der maximalen Zellfläche von $\chi = 80\,\mathrm{km}^2$.

Umgebungsvariable	*TSS*	*H*	*F*	*CSI*	*OLP*	*DIS*	Variablentrennwert	Einheit
T_{850hPa} (max)	0,29	0,61	0,32	0,02	0,73	0,27	286,9	K
DCI_{100hPa}	0,29	0,65	0,36	0,02	0,74	0,26	24,2	K
SCP	0,27	0,47	0,20	0,02	—	—	0,26	—
LI_{100hPa}	0,26	0,71	0,45	0,02	0,76	0,24	–0,9	K
SHIP	0,26	0,54	0,28	0,02	—	—	0,15	—
KO-Index	0,26	0,74	0,48	0,02	0,76	0,24	–4,6	K
SI	0,25	0,56	0,31	0,02	0,76	0,24	–0,9	K
0 °C-Grenze	0,24	0,71	0,47	0,02	0,78	0,22	3 306	m
$\theta_{ps,850hPa}$ (max)	0,24	0,68	0,44	0,02	0,78	0,22	325,2	K
T_{850hPa} (sdam)	0,22	0,66	0,44	0,02	—	—	0,24	K
MLS	0,22	0,57	0,35	0,02	0,79	0,21	9,5	$\mathrm{m\,s^{-1}}$
DLS	0,21	0,56	0,35	0,01	0,80	0,20	14,2	$\mathrm{m\,s^{-1}}$
$\overline{U}_{3-6km}$	0,20	0,48	0,28	0,01	0,80	0,20	15,5	$\mathrm{m\,s^{-1}}$
T_{2m}	0,20	0,65	0,45	0,02	0,80	0,20	291,1	K
500 hPa Geopot.	0,20	0,58	0,38	0,02	0,81	0,19	562,5	gpdam



(a) Unterscheidungsvermögen kurze / lange Lebensdauer



(b) Unterscheidungsvermögen kleine / große maximale Zellfläche

Abbildung 5.16: Darstellung verschiedener Gütemaße für das Unterscheidungsvermögen zwischen Zellobjekten unterschiedlicher Lebensdauer bzw. maximaler Zellfläche aus einer Statistik von 33 Umgebungsvariablen in Abhängigkeit vom Klassentrennwert (a) der Lebensdauer T_Z bzw. (b) der maximalen Zellfläche $A_{Z,max}$.

gezeigt). Mehr als 75 % der großen Zellobjekte haben größere Zellflächen zu diesen Zeitpunkten des Lebenszyklus, jedoch lediglich 10–15 % der kleineren Zellobjekte, die maximal 80 km^2 in ihrem Lebenszyklus groß werden. Dies bedeutet im Umkehrschluss, dass ein Zellobjekt bereits zu Beginn stark wachsen sollte, um eine große Fläche im Laufe des Lebenszyklus zu erreichen. Bereits in den Abbildungen 5.6 und 5.9 wurde ebenfalls deutlich, dass ein intensives anfängliches Wachstum auf eine große Zellfläche und zusätzlich auf eine lange Lebensdauer hindeutet.

Qualitativ ändern sich die gezeigten Ergebnisse bezüglich des Unterscheidungsvermögens der Umgebungsvariablen kaum für verschiedene Klassentrennwerte der Lebensdauer oder der maximalen Zellfläche ($\tau \in [15; 120]$ min bzw. $\chi \in [20; 90]$ km^2). Je niedriger dieser Klassentrennwert ist, desto geringer ist auch das Mittel oder der Median der TSS für die 33 untersuchten Variablen (Abbildung 5.16). Dieses Ergebnis lässt sich auch auf einen größeren Satz von Variablen erweitern. Der Anstieg der TSS

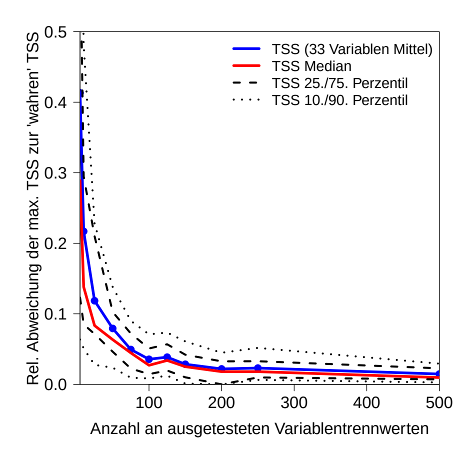


Abbildung 5.17: Relativer Unterschied der maximalen TSS für unterschiedliche Anzahlen von ausgetesteten Variablentrennwerten im Wertebereich der Umgebungsvariablen am Beispiel der 33 ausgewählten Umgebungsvariablen für einen Klassentrennwert der Lebensdauer von $\tau = 100$ min, bezüglich der maximalen TSS aus 1 000 ausgetesteten Trennwerten („wahre“ TSS).

lässt sich hauptsächlich durch einen Anstieg von H mit steigendem Klassentrennwert erklären, während F näherungsweise konstant bleibt. Der optimale Variablentrennwert ändert sich insgesamt nur wenig (nicht gezeigt).

Beispielsweise verschiebt sich für die DLS anschaulich gesprochen die Verteilung der Zellobjekte mit kurzer Lebensdauer (rote Verteilung in Abbildung 5.15a) bei steigendem Klassentrennwert kaum nach links hin zu kleineren DLS-Werten, während die Verteilung der Objekte mit langer Lebensdauer (blau) – deren Anzahl bei steigendem Klassentrennwert stetig abnimmt – merklich nach rechts wandert. Der optimale Variablentrennwert in Bezug auf die TSS variiert für die DLS nur um $\mathscr{O}(1\,\mathrm{m\,s^{-1}})$. Umgekehrt sinkt der CSI mit steigendem Klassentrennwert für die Lebensdauer oder die maximale Zellfläche aufgrund der Verschiebung der Anteile der Zellobjekte in den Gruppen. Die vielen Zellobjekte mit kurzer Lebensdauer rechts des Variablentrennwerts fallen hier sehr stark ins Gewicht.

Zuletzt sei kurz auf die Bestimmung der jeweils optimalen TSS (bzw. CSI) eingegangen, welche abhängig von der Anzahl der getesteten Variablentrennwerte ist. Je mehr Trennwerte der Algorithmus testet, umso rechenzeitintensiver gestaltet sich dies für eine große Anzahl von Umgebungsvariablen. Je weniger Trennwerte er testet, desto ungenauer wird die Schätzung der maximalen Scores in der Regel. Zur Untersuchung dieser Abhängigkeit wird für die ausgewählten 33 Umgebungsvariablen die Auswirkung unterschiedlich hoher Auflösungen der Variablentrennwerte auf die maximale TSS hinsichtlich der oben beschriebenen Unterscheidung von Zellobjekten mit kurzer und langer Lebensdauer betrachtet (gemittelt über alle 33 Umgebungsvariablen; Klassentrennwert der Lebensdauer: $\tau = 100\,\text{min}$; Abbildung 5.17). Die mittlere relative Abweichung zwischen den TSS-Werten mit 100 und 1 000 getesteten Trennwerten (Referenz) vom Minimum bis zum Maximum der Umgebungsvariablen beträgt lediglich 3,6 % ($\Delta TSS = \mathcal{O}(0{,}01)$). Für 1 000 Werte gilt die Annahme, dass die TSS für den optimalen Variablentrennwert der wahren TSS sehr nahe kommt. 100 Trennwerte sind demnach zur adäquaten Abschätzung der maximalen Scores ausreichend.

Erweiterter Parabelansatz

Eine Verfeinerung des Parabelansatzes nach Gleichung (5.1) für die zeitliche Entwicklung der Zellfläche über den Lebenszyklus (Kapitel 5.1.2) ist durch die Hinzunahme eines weiteren Scharparameters, einer Umgebungsvariablen u, möglich. Die minimale Zellfläche $A_{Z,min}^{(T_Z,u)} \approx A_{Z,min}^{(T_Z)} \approx \mu_A$ ist dabei von u nahezu unabhängig, während die Amplitude $\mathscr{A}^{(T_Z,u)}$ in der Regel eine Abhängigkeit aufweist.

Geht man weiterhin von einer linearen Abhängigkeit der Amplitude von der Lebensdauer aus, so lässt sich $\mathscr{A}^{(T_Z,u)} = c_A(u)T_Z$ setzen. Die Abhängigkeit des Koeffizienten c_A von der Umgebungsvariablen u ist dabei noch näher zu

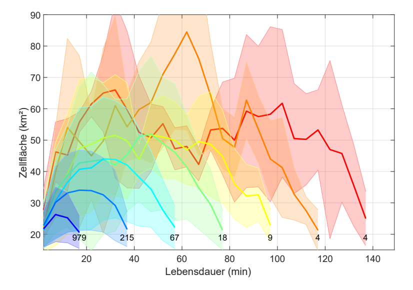
bestimmen. Der Ansatz lautet wie folgt:

$$A_Z^{(T_Z,u)}(t) = A_{Z,min}^{(T_Z)} + \mathscr{A}^{(T_Z,u)} - \frac{\mathscr{A}^{(T_Z,u)}}{(T_Z/2)^2}\left(t - \frac{T_Z}{2}\right)^2 \; . \tag{5.9}$$

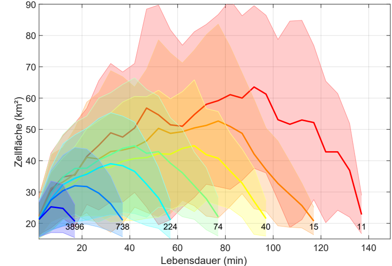
Exemplarisch sei dieser Ansatz mit dem *Lifted Index* ($\mathrm{LI}_{100\mathrm{hPa}}$ zum Zeitpunkt der ersten Detektion) vorgestellt. Je niedriger der $\mathrm{LI}_{100\mathrm{hPa}}$ ist, desto höher ist die latente Instabilität und damit die Möglichkeit für ein schnelles Wachstum einer Zelle durch freie Konvektion (s. o; Kapitel 2.3). Es lässt sich zunächst erkennen, dass die mittlere Amplitude $\mathscr{A}^{(T_Z,\mathrm{LI})}$ für niedrigere Werte des $\mathrm{LI}_{100\mathrm{hPa}}$ steigt (Abbildung 5.18). Es zeigt sich ein Unterschied von meist $2-10\,\mathrm{km}^2$ zwischen den mittleren Verläufen der Zellfläche von Objekten, die bei $\mathrm{LI}_{100\mathrm{hPa}} < -1$ K auftreten, und solchen, die bei $\mathrm{LI}_{100\mathrm{hPa}} \geq -1$ K auftreten. Die Wahl des Variablentrennwerts von -1 K basiert auf dem in Tabelle 5.2 gelisteten optimalen Variablentrennwert von $-0{,}9$ K. Darüber hinaus findet sich bei der Unterscheidung der Werte für die Lebensdauer ein optimaler Variablentrennwert von $-1{,}1$ K (nicht in Tabelle 5.1 gezeigt). Entsprechend wächst die Amplitude zu Beginn des Lebenszyklus bei Objekten, die bei $\mathrm{LI}_{100\mathrm{hPa}} < -1$ K auftreten, schneller an[5]. Es sei jedoch darauf hingewiesen, dass sich die Variationsbereiche (1σ-Intervalle) der Verläufe von Objekten gleicher Lebensdauer für $\mathrm{LI}_{100\mathrm{hPa}} < -1$ K und $\mathrm{LI}_{100\mathrm{hPa}} \geq -1$ K stark überlappen.

Die Überlappung der Variationsbereiche spiegelt sich in den Koeffizienten $c_A(\mathrm{LI})$ der linearen Regression (vgl. Kapitel 3.3.1) für verschiedene Wertebereiche des $\mathrm{LI}_{100\mathrm{hPa}}$ wider (Abbildung 5.19a). Zur Beschreibung der Abhängigkeit $c_A(\mathrm{LI})$ wird ein linearer Fit verwendet, welcher diese gut repräsentiert (Abbildung 5.19b). In diesen Fit gehen nur die Intervalle des $\mathrm{LI}_{100\mathrm{hPa}}$

[5] Es sei angemerkt, dass die Zellfläche zu Beginn des Lebenszyklus für den Strömungsfeldansatz (vgl. Kapitel 5.1.2) ebenfalls für niedrige Werte des $\mathrm{LI}_{100\mathrm{hPa}}$ (zum Zeitpunkt der ersten Detektion) schneller anwächst (nicht diskutiert).



(a) Mittlerer Verlauf der Zellfläche A_Z inklusive Variationsbereich für $\mathrm{LI}_{100\mathrm{hPa}} < -1\,\mathrm{K}$.



(b) Mittlerer Verlauf der Zellfläche A_Z inklusive Variationsbereich für $\mathrm{LI}_{100\mathrm{hPa}} \geq -1\,\mathrm{K}$.

Abbildung 5.18: Wie Abbildung 5.6b, nur zusammengefasst für Zellobjekte, die mit (a) niedrigeren bzw. (b) höheren Werten des $\mathrm{LI}_{100\mathrm{hPa}}$ assoziiert werden. Der Wert für den $\mathrm{LI}_{100\mathrm{hPa}}$ entspricht demjenigen zum Zeitpunkt der ersten Detektion der Zellobjekte.

ein, in denen mindestens neun unterschiedliche Werte der Lebensdauer vorkommen, die jeweils von mindestens zehn Zellobjekten erreicht wurden. Zur Bestimmung des jeweiligen c_A gehen daher nur wenige Zellobjekte mit einer langen Lebensdauer von mehr als 60 min ein. Analog zu Gleichung (5.2) erhält man folglich allgemein:

$$A_Z^{(T_Z,u)}(t) = \mu_A + 4c_A(u)t\left(1 - \frac{t}{T_Z}\right) . \tag{5.10}$$

Für $T_Z = 60\,\mathrm{min}$ ergibt sich nach Gleichung (5.2) beispielsweise eine maximale Zellfläche von $A_{Z,max}^{(60)} \approx 39\,\mathrm{km}^2$ (vgl. Abbildung 5.8a). In Abhängigkeit vom $\mathrm{LI}_{100\mathrm{hPa}}$ findet man mit $c_A(\mathrm{LI}) \approx (0{,}336\,-$

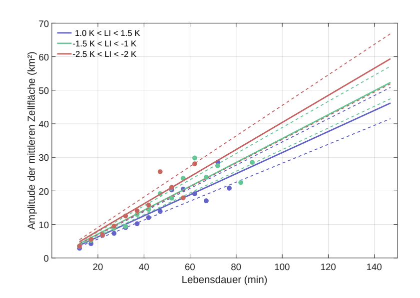
$0{,}024\,\mathrm{LI}_{100\mathrm{hPa}}\,\mathrm{K}^{-1})\,\mathrm{km}^2\,\mathrm{min}^{-1}$ über Gleichung (5.10) hingegen:

$$\begin{aligned}
\mathrm{LI}_{100\mathrm{hPa}} = 3\,\mathrm{K} &\implies A_{Z,max}^{(60,3)} \approx 37\,\mathrm{km}^2 \\
\mathrm{LI}_{100\mathrm{hPa}} = 1\,\mathrm{K} &\implies A_{Z,max}^{(60,1)} \approx 40\,\mathrm{km}^2 \\
\mathrm{LI}_{100\mathrm{hPa}} = -1\,\mathrm{K} &\implies A_{Z,max}^{(60,-1)} \approx 43\,\mathrm{km}^2 \\
\mathrm{LI}_{100\mathrm{hPa}} = -3\,\mathrm{K} &\implies A_{Z,max}^{(60,-3)} \approx 46\,\mathrm{km}^2 \\
\mathrm{LI}_{100\mathrm{hPa}} = -5\,\mathrm{K} &\implies A_{Z,max}^{(60,-5)} \approx 49\,\mathrm{km}^2\ .
\end{aligned} \tag{5.11}$$

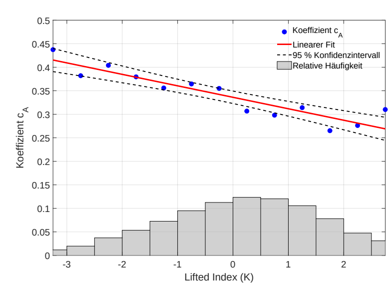
Erwartungsgemäß wächst mit steigender Instabilität die Zellfläche stärker an. Weitere Stabilitätsmaße wie beispielsweise die mitteltroposphärische *Lapse Rate* zeigen ähnliche Möglichkeiten zur Erweiterung des Parabelansatzes auf. Hingegen ist eine eindeutige lineare Abhängigkeit des Regressionskoeffizienten c_A von der Windscherung (z. B. DLS oder $\mathrm{SRH}_{0-\mathrm{xkm}}$) nicht gegeben, sodass eine Erweiterung des Parabelansatzes mit diesen Variablen nicht sinnvoll ist (nicht gezeigt).

5.3.2 Bivariate Analysen

Aufbauend auf den univariaten Ergebnissen folgt in diesem Abschnitt eine Diskussion der mittleren Lebensdauer sowie der mittleren maximalen Zellfläche der Zellobjekte in Abhängigkeit von einer Kombination aus zwei Umgebungsvariablen. Dazu spannen zwei Variablen einen gemeinsamen Raum auf, die nach Abbildung 5.13 möglichst wenig korreliert sind, um redundante Informationen zu vermeiden. Dies ist beispielsweise für die DLS und den $\mathrm{LI}_{100\mathrm{hPa}}$ der Fall, deren (Rang-)Korrelation $r_S = 0{,}15$ bzw. $r_P = 0{,}19$ beträgt. Die Differenz der mittleren Lebensdauer zwischen den Gruppen mit den niedrigsten und den höchsten Werten liegt maximal bei lediglich rund $10-15$ min (Abbildung 5.20), während die Standardabweichung im gleichen Bereich oder gar etwas höher liegt (Abbildung 5.21a). Der Unterschied der mittleren Lebensdauer bei einer Umgebungsvariablen alleine liegt jedoch mit



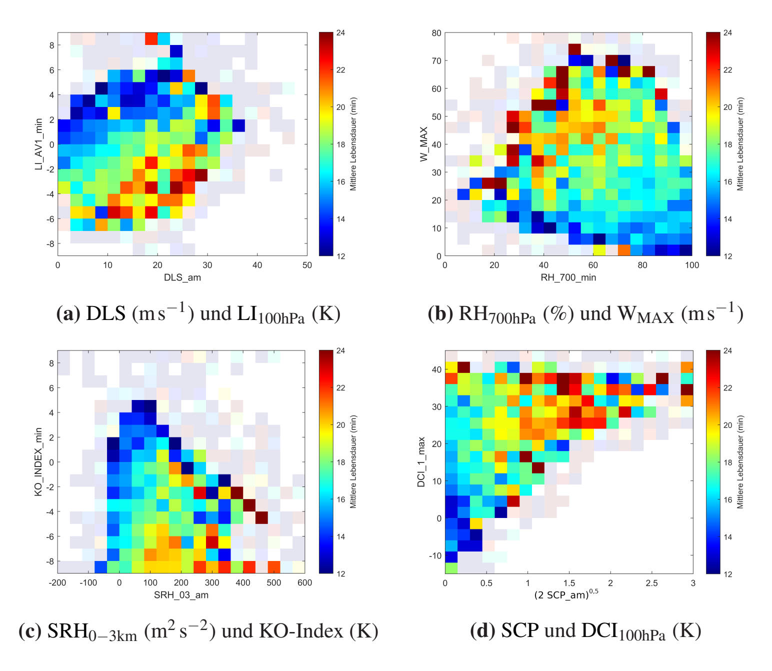
(a) Zur Illustration der Abhängigkeit des linearen Regressionskoeffizienten c_A von $\mathrm{LI}_{100\mathrm{hPa}}$ für die Amplitude $\mathscr{A}^{(T_Z,\mathrm{LI})}$.



(b) Zur Bestimmung der Abhängigkeit des Regressionskoeffizienten c_A von $\mathrm{LI}_{100\mathrm{hPa}}$. Die Regressionsgleichung lautet (*RMSE* $\approx$ 0,02 km^2 min^{-1}): $c_A \approx (0{,}336 - 0{,}024\,\mathrm{LI}_{100\mathrm{hPa}}\,\mathrm{K}^{-1})\,\mathrm{km}^2\,\mathrm{min}^{-1}$.

Abbildung 5.19: (a) Wie Abbildung 5.7a, nur für verschiedene Stichproben von Zellobjekten, die mit unterschiedlichen Wertebereichen des $\mathrm{LI}_{100\mathrm{hPa}}$ (Farbgebung) assoziiert werden. (b) Verteilung der linearen Regressionskoeffizienten c_A in Abhängigkeit vom $\mathrm{LI}_{100\mathrm{hPa}}$ (Intervalle von 0,5 K). Ein lineares Polynom approximiert wiederum diese Abhängigkeit (rote Linie). Die relative Häufigkeit von Zellobjekten, die in jedes Intervall fallen, ist als Histogramm hinzugefügt (0,1 = 10 % $\hat{=}$ 2 331 Objekten).

maximal 7 – 10 min noch niedriger (nicht gezeigt). Mit steigender mittlerer Lebensdauer steigt zudem im bivariaten Fall die Standardabweichung. Das Signal-Rausch-Verhältnis ist daher relativ gering. Dies könnte dadurch erklärt werden, dass an Tagen, an denen konvektive Zellen mit einer längeren Lebensdauer aufgrund besonders konvektionsförderlicher Umgebungsbedingungen auftreten, sehr häufig weitere Zellen auftreten, die eine weniger lange Lebensdauer haben. Dennoch steigt die mittlere Lebensdauer erwartungsgemäß für die Kombination aus hoher Windscherung und hoher Instabilität (niedrigem $\mathrm{LI}_{100\mathrm{hPa}}$) an. Des Weiteren befinden sich mehr als 75 % der langlebigen Zellen mit mehr als 90 Minuten Lebensdauer in denjenigen 50 % der Gruppen, welche die höchste mittlere Lebensdauer aufweisen (nicht gezeigt).



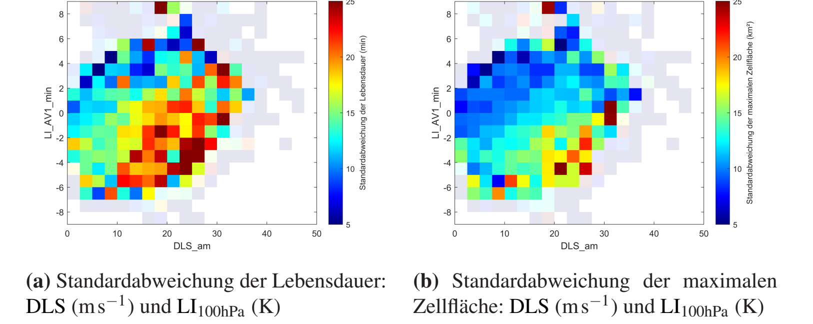
(a) DLS ($m\,s^{-1}$) und LI_{100hPa} (K)

(b) RH_{700hPa} (%) und W_{MAX} ($m\,s^{-1}$)

(c) SRH_{0-3km} ($m^2\,s^{-2}$) und KO-Index (K)

(d) SCP und DCI_{100hPa} (K)

Abbildung 5.20: Mittlere Lebensdauer aller 38 553 Zellobjekte gruppiert nach je zwei Umgebungsvariablen in je 20 Intervallen (zum Zeitpunkt der ersten Detektion). Gruppen, in denen weniger als zehn Zellobjekte zu finden sind, sind transparent dargestellt.



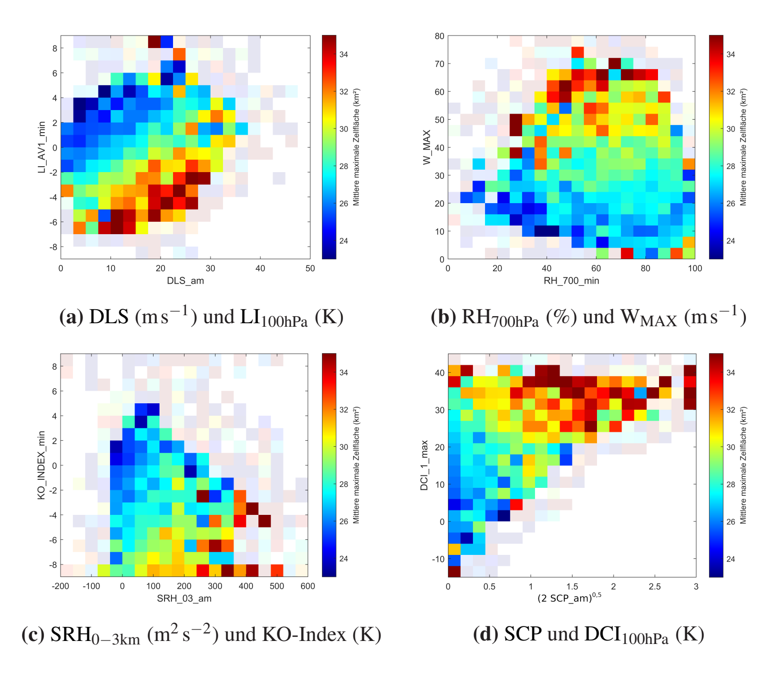
(a) Standardabweichung der Lebensdauer: DLS ($m\,s^{-1}$) und LI_{100hPa} (K)

(b) Standardabweichung der maximalen Zellfläche: DLS ($m\,s^{-1}$) und LI_{100hPa} (K)

Abbildung 5.21: Wie Abbildung 5.20, nur für die Standardabweichung (a) der Lebensdauer und (b) der maximalen Fläche aller Zellobjekte.

Eine geringe mitteltroposphärische Feuchte RH_{700hPa} und ein hoher Gehalt an $CAPE_{MU}$ (bzw. eine hohe W_{MAX}) bewirken ebenfalls eine längere Lebensdauer, wobei für hohe Feuchtewerte das Maß an konvektiv verfügbarer Energie weniger wichtig ist als für niedrige Feuchtewerte (Abbildung 5.20b). Niedrige Werte des KO-Index alleine reichen nicht aus, um im Mittel eine längere Lebensdauer hervorzubringen. Zusätzlich muss die SRH_{0-3km} deutlich größer als $0\,m^2\,s^{-2}$ sein (Abbildung 5.20c). Die Lebensdauer steigt ebenfalls, wenn sowohl der DCI_{100hPa} als auch der SCP hohe Werte erreichen (Abbildung 5.20d).
Qualitativ sehr ähnliche Ergebnisse findet man bei der Untersuchung der maximalen Zellfläche (Abbildung 5.22). Der Unterschied zwischen den verschiedenen Gruppen beträgt etwa $10-15\,km^2$ und die Verteilung ist mit derjenigen der Werte für die Lebensdauer annähernd deckungsgleich. Dies stimmt mit den Resultaten aus dem Parabelansatz und dem Strömungsfeldansatz überein, die auf einen klaren Zusammenhang zwischen den beiden Zellattributen hinweisen. Die Standardabweichung ist für die maximale Zellfläche mit meist weniger als $15\,km^2$ ebenfalls hoch (Abbildung 5.21b).

Ein etwas deutlicheres Signal für beide Zellattribute zeigt sich erst, wenn die Analyse Zellobjekte mit sehr niedrigen Werten für die Lebensdauer von weniger als z. B. 30 min nicht berücksichtigt (Abbildung 5.23). Die Differenz der mittleren Lebensdauer zwischen den Gruppen mit den niedrigsten und den höchsten Werten verdoppelt sich dadurch etwa, was der höheren relativen Häufigkeit von Zellobjekten mit einer eher längeren Lebensdauer in den einzelnen Gruppen geschuldet ist. Wenn sich demnach Zellobjekte unter gegebenen Umgebungsbedingungen nicht innerhalb einer halben Stunde wieder auflösen, erreichen sie bei hoher Instabilität und gleichzeitig hoher Windscherung im Mittel eine um $20-25$ min längere Lebensdauer und werden etwa $20-25\,km^2$ größer als bei geringer Instabilität und Windscherung. Die DLS hat einen stärkeren Einfluss auf die Lebensdauer (Abbildung 5.23a), während



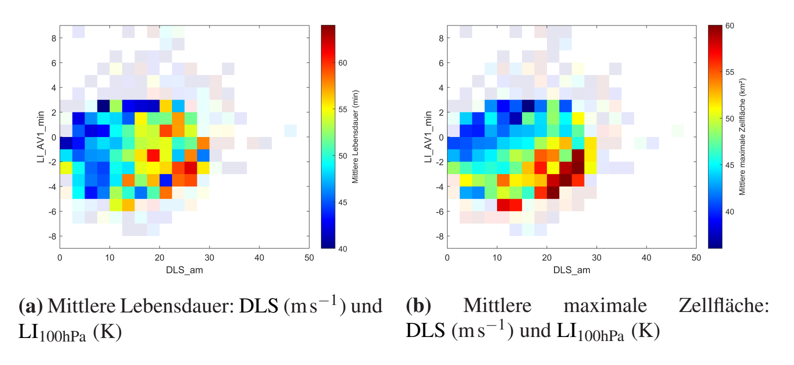
(a) DLS ($\mathrm{m\,s^{-1}}$) und $\mathrm{LI_{100hPa}}$ (K)

(b) $\mathrm{RH_{700hPa}}$ (%) und $\mathrm{W_{MAX}}$ ($\mathrm{m\,s^{-1}}$)

(c) $\mathrm{SRH_{0-3km}}$ ($\mathrm{m^2\,s^{-2}}$) und KO-Index (K)

(d) SCP und $\mathrm{DCI_{100hPa}}$ (K)

Abbildung 5.22: Wie Abbildung 5.20, nur für die mittlere maximale Fläche aller Zellobjekte.



(a) Mittlere Lebensdauer: DLS ($\mathrm{m\,s^{-1}}$) und $\mathrm{LI_{100hPa}}$ (K)

(b) Mittlere maximale Zellfläche: DLS ($\mathrm{m\,s^{-1}}$) und $\mathrm{LI_{100hPa}}$ (K)

Abbildung 5.23: Wie Abbildung 5.20, nur für (a) die mittlere Lebensdauer und (b) die mittlere maximale Fläche aller Zellobjekte, die eine Lebensdauer von mindestens 30 min haben.

der LI_{100hPa} stärker die maximale Zellfläche beeinflusst (Abbildung 5.23b), was sich mit den Ergebnissen aus den univariaten Analysen in Kapitel 5.3.1 deckt (Tabellen 5.1 und 5.2).

Auf die Darstellung und Diskussion von kombinierten Abhängigkeiten von mehr als zwei Umgebungsvariablen wird an dieser Stelle verzichtet. Erste Untersuchungen mit Kombinationen von drei bis sechs Variablen im Rahmen der vorliegenden Arbeit zeigen, dass mit einer Erhöhung der Dimensionalität des Variablenraums auf drei bis sechs Dimensionen kaum Untermannigfaltigkeiten auffallen, die gleichzeitig (i) genügend Zellobjekte für eine statistische Untersuchung beinhalten und (ii) bedeutend stärkere Signale hinsichtlich der beobachteten Lebensdauer oder maximalen Zellfläche zeigen als die bislang diskutierten Ergebnisse. Die Untersuchungen im nachfolgenden Kapitel 6 werden daher unter anderem der Fragestellung nachgehen, inwiefern die Kombination von mehreren Variablen in Vorhersageverfahren für die Abschätzung der Lebensdauer und der maximalen Zellfläche hilfreich sein kann.

6 Vorhersageverfahren: Entwicklung und Evaluation

Die Analysen aus Kapitel 5 haben einige statistische Zusammenhänge zwischen verschiedenen Attributen konvektiver Zellen, wie der Lebensdauer und der maximalen Zellfläche, und verschiedenen Umgebungsvariablen sowie der anfänglichen Zellentwicklung aufgezeigt. In der Folge stellen sich nun mehrere Fragen, beispielsweise wie diese Ergebnisse für das *Nowcasting* konvektiver Zellen gewinnbringend eingebracht werden könnten. Dabei ist von besonderer Relevanz, ob aufgrund der großen Dominanz der Häufigkeiten von Zellobjekten mit kurzer Lebensdauer (bzw. kleiner Zellfläche) auf einem historischen Datensatz beruhende Vorhersageverfahren überhaupt zwischen solchen und Zellobjekten mit langer Lebensdauer (bzw. großer Zellfläche) unterscheiden können. Als Erweiterung der uni- und bivariaten Analysen liegt zudem die Frage nahe, ob durch die Anwendung von multivariaten Verfahren möglicherweise Kombinationen von mehreren atmosphärischen Umgebungsvariablen und/oder der anfänglichen Entwicklung der Zellobjekte identifiziert werden können, die ein gutes Unterscheidungsvermögen der zu erwartenden konvektiven Entwicklungen besitzen.
Diese Fragen motivierten zu einer eingehenden Beschäftigung mit multivariaten Verfahren. Drei verschiedene Verfahren der Statistik und des maschinellen Lernens, die sich für die vorliegende Arbeit als potentiell geeignet herausstellten und in den nachfolgenden Kapiteln Verwendung finden werden, wurden in den Kapiteln 3.3 und 3.4 bereits ausführlich beschrieben: die logistische Regression, der *Random Forest* und ein nicht-linearer Polynomansatz (mit der linearen Regression als eine Art Spezialfall). Diese Verfahren können auf

der Basis des in Kapitel 5 analysierten kombinierten Datensatzes eine deterministische oder probabilistische Vorhersage der Lebensdauer oder der maximalen Fläche eines Zellobjekts treffen. Die Informationen haben dabei das Potential, das probabilistische Lebenszyklusmodell des in Entwicklung befindlichen Zellverfolgungsalgorithmus KONRAD3D (DWD) zu verbessern, indem die Schwankungsbreite des Modellensembles insbesondere zu Beginn der Zellentwicklung besser eingegrenzt werden kann (vgl. Kapitel 5.1.2; Feger et al., 2019; Werner, 2020; Wapler, 2021). Zudem könnte es für weitere Fragestellungen und Anwendungen des *Nowcastings* nützlich sein, bestimmte Klassen der Zellattribute frühzeitig abschätzen zu können (z. B. kurze/lange Lebensdauer). Kategorische oder kontinuierliche Abschätzungen der Lebensdauer und der Zellfläche sind zudem zur besseren zeitlichen bzw. räumlichen Spezifizierung von Warnungen von Relevanz.
Die oben erwähnten multivariaten Klassifikations- und Regressionsverfahren werden im Folgenden in verschiedenen Modellstudien als Vorhersageverfahren angewendet, die unterschiedliche Kombinationen von Umgebungsvariablen und/oder Zellattributen als Prädiktoren verwenden (vgl. Abbildung 6.1). Kapitel 6.1 beschreibt zunächst die notwendige Datenvorbehandlung, welche vor der Anwendung eines Vorhersageverfahrens erfolgen muss, und rekapituliert einige Besonderheiten des kombinierten Datensatzes, die unter anderem in der Evaluation der Verfahren berücksichtigt werden. Anschließend folgt mit den zwei Klassifikationsverfahren logistische Regression und *Random Forest* die Diskussion einer ersten Modellstudie mit zwei Prädiktoren, anhand derer viele verschiedene Aspekte solcher Verfahren und der entsprechenden Evaluation detailliert erläutert werden (Kapitel 6.2). Die Kapitel 6.3 und 6.4 fassen die Ergebnisse aus vielen weiteren Modellstudien zusammen und analysieren und vergleichen das Potential der Verfahren für die Abschätzung der Lebensdauer und der maximalen Zellfläche. Darüber hinaus wird die (statistische) relative Wichtigkeit der Umgebungsvariablen und Zellattribute als Prädiktoren für die Vorhersage diskutiert.

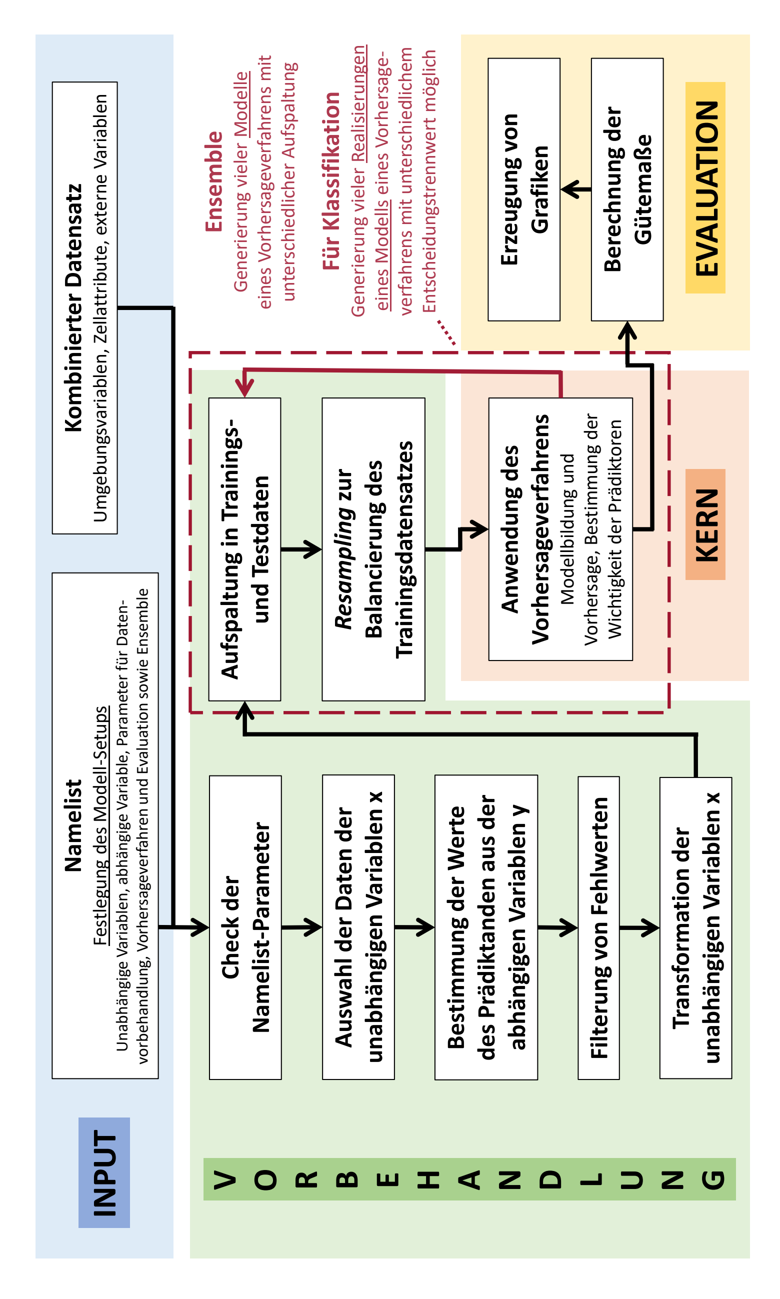


Abbildung 6.1: Schematische Vorgehensweise in den verschiedenen Modellstudien zur Abschätzung der zu erwartenden Lebensdauer und maximalen Fläche von konvektiven Zellen mittels unterschiedlicher Vorhersageverfahren.

6.1 Besonderheiten in der Datenvorbehandlung und der Evaluation

Für die Modellstudien, die in der vorliegenden Arbeit vorgestellt werden, müssen folgende Eigenschaften und Besonderheiten berücksichtigt werden:

(i) Die unabhängigen Variablen sind teilweise stark korreliert, insbesondere einige Umgebungsvariablen (vgl. Kapitel 5.2.2).

(ii) Die unabhängigen Variablen variieren auf sehr unterschiedlichen Skalen und sind meist nicht normalverteilt (vgl. Kapitel 5.2.1).

(iii) Der Datensatz weist ein starkes Ungleichgewicht in der Verteilung der Werte der ausgewählten abhängigen Variablen auf: der Lebensdauer und der maximalen Zellfläche (vgl. Kapitel 5.1.1).

(iv) In Modellstudien, die mit einem Ensembleansatz durchgeführt werden, werden keine einheitlichen Trainings- und Testdatensätze für die Modellbildung bzw. die Vorhersagen der Mitglieder innerhalb des Ensembles verwendet, um die Robustheit der Modelle bezüglich des Datensatzes möglichst genau abzuschätzen (vgl. Kapitel 3.6.1).

Die genannten Punkte führen dazu, dass sowohl in der Vorbehandlung des Datensatzes vor der Modellbildung als auch in der Evaluation nach der Anwendung der Modelle einige weitere Schritte vorgenommen werden müssen, um eine adäquate Durchführung der Modellstudien und eine entsprechende Interpretation der Ergebnisse zu ermöglichen.

6.1.1 Datenvorbehandlung zur Anwendung der Vorhersageverfahren

Auswahl der Prädiktoren

Wie beispielsweise in Wilks (2006) erläutert, ist ein Regressionsverfahren mit beliebigen N_{po} Prädiktoren in der Lage, die Residuen eines Trainingsdatensatzes der Größe $N_{Tr} = N_{po} + 1$ verschwinden zu lassen, d. h. einen perfekten Fit zu generieren. Dies bedeutet jedoch nicht, dass ein perfekt trainiertes Modell ebenso gute Vorhersagen für einen unabhängigen Testdatensatz trifft. Im Gegenteil: Ein solches Modell ist oftmals viel zu sehr an den zugrundeliegenden Trainingsdatensatz angepasst (*Overfitting*). Daher ist es sinnvoll, bereits im Voraus eine Auswahl von potentiell relevanten unabhängigen Variablen $\mathbf{x}$ zu treffen, die physikalisch mit der gewählten abhängigen Variablen y in Verbindung stehen. Je größer der zugrunde liegende Datensatz ist, desto mehr (voneinander unabhängige) Prädiktoren kann ein Vorhersageverfahren in der Regel sinnvoll miteinbeziehen. In der Praxis helfen jedoch für eine (nicht-)lineare Regression mehr als etwa zwölf Prädiktoren selten, die Vorhersagen noch weiter zu verbessern (Glahn, 1985). Der *Random Forest* weist hingegen seine Stärken eher mit einer großen Anzahl von Prädiktoren auf, vorausgesetzt der Anteil von relevanten Prädiktoren mit einem positiven Einfluss auf die Vorhersage ist nicht zu klein (z. B. Hastie et al., 2009). Der Grund dafür ist die stärkere Dekorrelation der Entscheidungsbäume aufgrund der größeren Verfügbarkeit von potentiellen Kandidaten für das Splitting der Äste der Entscheidungsbäume (vgl. Kapitel 3.4.3).

Ein finales Modell mit einem optimal angepassten Satz von Prädiktoren und Modellparametern zu bestimmen, ist nicht Ziel dieser Arbeit. Vielmehr steht die Analyse des Vorhersagevermögens der drei verschiedenen multivariaten Verfahren sowie der Vergleich ihrer Ergebnisse im Vordergrund. Daher beruht die Auswahl der Prädiktoren für die Modellstudien in der vorliegenden Arbeit nicht auf einer statistisch systematischen Vorgehensweise wie der schrittweisen Regression (Wilks, 2006). Um verschiedene Vorhersageverfahren mit

demselben Satz von unabhängigen Variablen miteinander vergleichen zu können, erfolgt die Auswahl der Prädiktoren in der vorliegenden Arbeit stattdessen auf Basis der Analysen in Kapitel 5. Sowohl Umgebungsvariablen als auch Zellattribute finden dabei Verwendung. Detaillierte Erläuterungen sind jeweils zu Beginn der Beschreibung der Modellstudien angegeben.

Filterung von Fehlwerten

Die Aufteilung des kombinierten Datensatzes in Trainings- und Testdaten ist für die unabhängige Evaluation der Vorhersageverfahren essentiell. Zuvor sortiert der Algorithmus, dessen Struktur sich an der Darstellung in Abbildung 6.1 orientiert, bereits alle Zellobjekte aus dem Datensatz aus, denen der Wert von mindestens einem Prädiktor fehlt (vgl. Kapitel 4.3.4). Beispielsweise kann nicht bei allen Zellobjekten das NFK berechnet und zugewiesen werden. In Modellstudien, bei denen Zellattribute zu bestimmten Zeitpunkten des Lebenszyklus (z. B. die Zellfläche 15 min nach der ersten Detektion) als Prädiktoren dienen, sortiert der Algorithmus alle Zellobjekte aus, deren Lebensdauer geringer ist als der betrachtete Zeitpunkt.

Transformation der Werte der Prädiktoren

Wie in Kapitel 3.5.1 geschildert kann eine mathematische Transformation der Werte der Prädiktoren vor der Anwendung eines statistischen Verfahrens nützlich sein, dessen Aussagekraft zu steigern. Czernecki et al. (2019) wendeten eine Kombination von Yeo-Johnson- und z-Transformation beispielsweise für die Vorhersage von Hagel mit Hilfe eines *Random Forests* an. Diese Kombination findet auch in den Modellstudien ab Kapitel 6.2 Anwendung. Der Einfluss der Transformationen auf die Vorhersage der Zellattribute und deren Güte sowie auf die Wichtigkeit der Prädiktoren wird für die erste Modellstudie aus Kapitel 6.2 in Anhang B diskutiert. Eine Transformation der Werte der abhängigen Variablen y, die für Regressionsverfahren relevant sein kann, stellte sich im Rahmen der vorliegenden Arbeit nicht als vorteilhaft

heraus. Eine Erhöhung der Schärfe und Verbesserung der Güte der Vorhersage, insbesondere für Regressionsverfahren, lässt sich durch das *Resampling* erreichen (s. u.; vgl. Kapitel 3.5.2).

Aufspaltung in Trainings- und Testdaten

Nach der Transformation der Werte der Prädiktoren erfolgt die Aufspaltung des gesamten Datensatzes in potentielle Trainingsdaten und Testdaten dergestalt, dass in beiden Datensätzen die ursprüngliche Verteilung der Werte der Prädiktanden erhalten bleibt. Den Anteil der potentiellen Trainingsdaten, die einen Pool von Zellobjekten für die Auswahl der finalen Trainingsdaten darstellen, gibt f_{Tr} vor (s. Kapitel 6.2.1). Für die Aufspaltung wird bei Klassifikationsverfahren in der vorliegenden Arbeit zwischen (a) Studien mit variablem Entscheidungstrennwert μ (verschiedene Realisierungen eines Modells; vgl. Kapitel 3.3.2 und 3.4.3) und (b) Studien mit festem Entscheidungstrennwert mit einem Modellensemble (vgl. Kapitel 3.6.1) unterschieden. Bei Regressionsverfahren finden ausschließlich Ensemblestudien statt. Im Fall (a) erhält jede Realisierung – sofern nicht anders bei den Auswertungen beschrieben – dieselben finalen Trainings- und Testdaten. Im Fall (b) erfolgt eine separate Aufteilung der potentiellen Trainingsdaten und Testdaten für jedes Ensemblemitglied.

Die Testdaten werden jeweils beiseite gelegt, während alle potentiellen Trainingsdaten selbst oder ein Satz aus N_{Tr} pseudo-zufällig und mit Zurücklegen gezogenen Zellobjekten als finaler Trainingsdatensatz dienen. Letzteres nennt sich *Bootstrapping* und zählt zu den *Resampling*-Methoden. Im Fall (a) kann man das *Bootstrapping* nutzen, um bei festem Testdatensatz durch die Verwendung unterschiedlicher Startwerte (*Seeds* = Samen) des Pseudo-Zufallszahlengenerators unterschiedliche Modelle zu bilden. Im Fall (b) unterscheiden sich die potentiellen Trainingsdaten ohnehin zwischen den einzelnen Mitgliedern des Ensembles. Ein *Bootstrapping* ist hier daher nur notwendig, wenn die erforderliche Größe des Trainingsdatensatzes die Anzahl

von potentiellen Trainingsdaten übersteigt. In allen Fällen folgen die mit den Trainingsdaten assoziierten Zellattribute weiterhin in sehr guter Näherung der ursprünglichen Verteilung des gesamten Datensatzes. Da es hier nicht das Ziel ist, ein Modell mit einem optimal angepassten Satz von Prädiktoren und Modellparametern zu bestimmen (s. o.), findet neben dem Ensembleansatz generell keine Kreuzvalidierung mittels eines dritten, von Trainings- und Testdaten unabhängigen Validierungsdatensatzes statt.

Resampling zur Balancierung des Trainingsdatensatzes

Wie in Kapitel 3.5.2 erläutert wurde, können *Resampling*-Methoden Probleme verringern, die ein unbalancierter Datensatz mit sich bringt, in dem die Werte der abhängigen Variablen sehr ungleich verteilt sind. In einigen Modellstudien in den Kapiteln 6.3 und 6.4 findet eine Kombination von *Undersampling* und *Oversampling* Anwendung, da sich diese in vielen Fällen als vorteilhaft herausgestellt hat: Der *Random Forest* kann nur dank des *Resamplings* in bestimmten Setups effizient verwendet werden (s. Kapitel 6.3.1 und 6.4.1). Die Auswirkungen verschiedener *Resampling*-Methoden auf die Vorhersagegüte der zwei Klassifikationsverfahren sind eher gering (Anhang B). Die Regressionsverfahren hingegen profitieren vom *Resampling* durch eine Erhöhung der Schärfe und eine Verbesserung des Vorhersage-Bias (s. Kapitel 6.3.2 und 6.4.2).

6.1.2 Bedingte Evaluation und spezielle Ensembleevaluation

Zur differenzierten Interpretation der Ergebnisse der Modellstudien folgt ergänzend zu der allgemeinen Beschreibung von Evaluationsmaßen (Kapitel 3.6) eine kurze Zusammenstellung der besonderen Aspekte der Evaluation, die speziell für die Modellstudien der vorliegenden Arbeit relevant sind. Sind Ereignisse im gesamten Datensatz und somit im Testdatensatz deutlich seltener vertreten als Nicht-Ereignisse, gilt für die Kontingenztabelle der binären Vorhersage eines Klassifikationsverfahrens (Tabelle 3.1) allgemein

$b+d \gg a+c$ und meist auch $b,d \gg a,c$. Aus diesem Grund ist folglich der *Proportion Correct* (*PC*; vgl. Tabelle 3.2) meist durch $PC \approx dN_{Te}^{\prime -1}$ gegeben, da $d \gg a$ ist. Bei der Interpretation dieses Gütemaßes ist folglich besondere Vorsicht geboten. Andere Gütemaße wie die Trefferrate H, die Fehlalarmrate F und die *True Skill Statistic* TSS sind unabhängig von der Verteilung der Werte einer binären abhängigen Variablen. Im Fall des konstruierten binären Prädiktanden Lebensdauer (kurz/lang) dominieren jedoch beispielsweise die am häufigsten erreichten Werte für die Lebensdauer die Werte von H und F. Bei den Zellobjekten mit kurzer Lebensdauer dominieren die Objekte mit sehr kurzer Lebensdauer (7, 12, 17 min) und bei den Zellobjekten mit langer Lebensdauer diejenigen mit einer Lebensdauer nahe des Klassentrennwerts (abhängig von der Wahl des Übergangsbereichs; vgl. Kapitel 3.6.1).

Eine mögliche Ergänzung zur Berücksichtigung dieses Ungleichgewichts der Stichprobe ist die Formulierung balancierter Gütemaße, welche speziell für die Evaluation von Ensemblevorhersagen verwendet werden (s. u.). Prinzipiell wäre auch a priori ein *Resampling* des gesamten Datensatzes oder des jeweiligen Testdatensatzes möglich, um die Verteilung der Werte der abhängigen Variablen auszugleichen (vgl. Kapitel 3.5.2 und 6.1.1). In der vorliegenden Arbeit entspricht die Verteilung im Testdatensatz jedoch stets in etwa der Verteilung im originalen gesamten Datensatz, sodass balancierte Gütemaße als Ergänzung zu den standardmäßigen Gütemaßen dienen. Generell hilft die Betrachtung mehrerer Gütemaße, verschiedene Aspekte der Vorhersagen näher zu beleuchten (vgl. Tabelle 3.2; Wilks, 2006), wie auch schon Doswell et al. (1990) in einer Studie zur Güte von Tornado-Vorhersagen empfahlen – ebenfalls eine Fragestellung, in der ein Ungleichgewicht des Datensatzes präsent ist.

Bedingte Evaluation

Die grundlegende Idee hinter der Formulierung eines balancierten Gütemaßes ist es, sich den gesamten Wertebereich der abhängigen Variablen y, die hier im Fall von Klassifikationsverfahren in einen binären Prädiktanden überführt wird, zu Nutze zu machen. Auch wenn es sich um eine binäre Vorhersage handelt, ist es möglich, die Güte der Vorhersagen für endlich viele (N_I) Intervalle I des Wertebereichs von y separat zu quantifizieren. Beispielsweise kann die Lebensdauer der zeitlichen Auflösung der Radarmessungen entsprechend in Intervalle von 5 min aufgeteilt werden. Diese Betrachtungsweise bezeichnet man als bedingte Evaluation, welche in den Kapiteln 6.2 bis 6.4 zusätzlich mit Grafiken veranschaulicht wird, welche die Ensemblevorhersagen den jeweiligen Beobachtungen der abhängigen Variablen y gegenüberstellen. Dies geschieht mittels einer auf die jeweilige Problemstellung angepassten Darstellung von bedingten Quantil-Plots (*Conditional Quantile Plots*) nach Murphy et al. (1989) basierend auf der sogenannten *Likelihood-Base Rate Factorization* sowie der *Calibration-Refinement Factorization* (Murphy und Winkler, 1987; Wilks, 2016).

Ein balanciertes Gütemaß für binäre Klassifikationsverfahren

Ein neu eingeführtes Maß für die balancierte Genauigkeit (*Accuracy*; *ACC*) einer Ensemblevorhersage für Klassifikationsverfahren stellt

$$ACC = \frac{1}{N_I} \sum_{i=1}^{N_I} acc_i \tag{6.1}$$

mit

$$acc_i = \frac{1}{N_{I_i}} \sum_{\{j \mid y_j^{(obs)} \in I_i\}} \left[\left(1 - \hat{y}_j^{(ens)}\right) \delta_{y_j,0} + \hat{y}_j^{(ens)} \delta_{y_j,1} \right] \tag{6.2}$$

dar (vgl. Kapitel 3.6.1). Darin steht $y_j^{(obs)}$ zur Abgrenzung gegenüber dem binären Prädiktanden y_j für den beobachteten Wert der nicht-binären, abhängigen Variablen y. Die spezifische balancierte Genauigkeit acc_i gibt folglich die mittlere Wahrscheinlichkeit für eine korrekte Vorhersage eines Zellobjekts innerhalb des i-ten Intervalls mit N_{I_i} Beobachtungen basierend auf einem Ensemble an. Das balancierte Gütemaß ACC entspricht darüber hinaus der über alle N_I Intervalle gemittelten mittleren Wahrscheinlichkeit für eine korrekte Vorhersage eines Zellobjekts basierend auf einem Ensemble. Damit folgt, dass $ACC \in [0;1]$ ist mit $ACC_{opt} = 1$ als optimalem Wert. Mit dieser Formulierung eines Gütemaßes werden alle Intervalle gleichermaßen berücksichtigt – unabhängig von der Anzahl von Zellobjekten, die in die jeweiligen Intervalle fallen. Wird zur Separation der beiden Klassen des Prädiktanden bei der Evaluation ein Übergangsbereich um den Klassentrennwert gewählt, so gehen in Gleichung (6.1) nur diejenigen Intervalle ein, die außerhalb dieses Übergangsbereichs liegen (vgl. Kapitel 3.6.1). Damit sehr wenige Zellobjekte mit extrem selten aufgetretenen Werten der abhängigen Variablen die ACC nicht zu stark dominieren, sollte zudem eine Mindestanzahl von Objekten mit gleichen Werten in einem Intervall vorliegen, damit die acc dieses Intervalls in die Berechnung der ACC eingehen darf.
Für ein Ensemble, bei dem jedem Mitglied derselbe Testdatensatz zugrundeliegt, entspricht die ACC dem Mittel der über das Ensemble gemittelten PC-Werte in den jeweiligen Intervallen, d. h. es gilt

$$ACC = PC_{bal} = \frac{1}{N_I} \sum_{i=1}^{N_I} PC_i^{(ens)} \tag{6.3}$$

mit dem Ensemblemittel der Scores der einzelnen Mitglieder $PC_i^{(q)}$ in den jeweiligen Intervallen

$$PC_i^{(ens)} = \frac{1}{N_{ens}} \sum_{q=1}^{N_{ens}} PC_i^{(q)} \; . \tag{6.4}$$

Für jedes Intervall wird demnach für jedes Ensemblemitglied eine eigene Kontingenztabelle (Tabelle 3.1) erstellt. Gehört ein Intervall I_i zur Klasse der Nicht-Ereignisse, folgt, dass $a_i^{(q)} = c_i^{(q)} = 0$ und $PC_i^{(q)} = d_i^{(q)} N_{I_i}^{-1}$ mit $N_{I_i} = b_i^{(q)} + d_i^{(q)}$ ist. Umgekehrt ist für ein Intervall I_i, das zur Klasse der Ereignisse gehört, $PC_i^{(q)} = a_i^{(q)} N_{I_i}^{-1}$ mit $N_{I_i} = a_i^{(q)} + c_i^{(q)}$. Für ein Ensemble, bei dem nicht jedem Mitglied derselbe Testdatensatz zugrundeliegt, ist $ACC \neq PC_{bal}$. Da dies in den folgenden Modellstudien der Fall sein wird (vgl. Kapitel 6.1.1), wird die balancierte Genauigkeit ACC über die Gleichungen (6.1) und (6.2) berechnet.

Ein balanciertes Gütemaß für Regressionsverfahren

Ein neu eingeführtes Maß für die balancierte Genauigkeit einer Ensemblevorhersage für Regressionsverfahren stellt in Analogie zum *Root Mean Squared Error* (*RMSE*) die Wurzel des balancierten, mittleren quadratischen Fehlers bzw. des *Balanced Root Mean Squared Errors* (*BRMSE*) dar. Dieser wird in den Modellstudien mit Regressionsverfahren über

$$BRMSE = \frac{1}{N_I} \sum_{i=1}^{N_I} me_i^2 \tag{6.5}$$

mit

$$me_i = \left\langle \frac{1}{N'_{ens}} \sum_{q=1}^{N'_{ens}} \left(y_j - \hat{y}_j^{(q)} \right) \right\rangle_{\{j \mid y_j \in I_i\}} \tag{6.6}$$

berechnet (vgl. Kapitel 3.3.1). Zunächst wird demnach die über das Ensemble gemittelte Abweichung der Vorhersagen der Mitglieder für das j-te Zellobjekt vom beobachteten Wert bestimmt. N'_{ens} steht für die Anzahl an Mitgliedern, bei denen das j-te Zellobjekt im Testdatensatz vorhanden ist (vgl. Kapitel 3.6.1). Dies erfolgt für alle Objekte innerhalb des Intervalls I_i. Im Anschluss ergibt sich der dortige spezifische mittlere Fehler me_i als der Median $\langle \cdot \rangle$ der gemittelten Abweichungen aller Zellobjekte, für die $y_j \in I_i$ gilt. Der *BRMSE* ist schließlich das arithmetische Mittel der Quadrate von me

über alle Intervalle. Damit folgt, dass $BRMSE \in [0\,;\infty]$ ist mit $BRMSE_{opt} = 0$ als optimalem Wert. Für jedes Intervall geht demnach in die Berechnung des quadratischen Fehlers der Wert eines Stellvertreters ein, in dieser Formulierung der Median der Ensemblemittelwerte. Mit dieser Formulierung eines Gütemaßes werden wie für die *ACC* alle Intervalle gleichermaßen berücksichtigt – unabhängig von der Anzahl von Zellobjekten, die in die jeweiligen Intervalle fallen. Damit sehr wenige Zellobjekte mit extrem selten aufgetretenen Werten der abhängigen Variablen den *BRMSE* nicht zu stark dominieren, sollte hier ebenfalls eine Mindestanzahl von Objekten mit gleichen Werten in einem Intervall vorliegen, damit der *me* dieses Intervalls in die Berechnung des *BRMSE* eingehen darf.

Spezielle probabilistische Gütemaße

Die allgemeine Formulierung für den Brier Score (*BS*) in Gleichung (3.77) kann modifiziert werden, um für jede Klasse des Prädiktanden separat einen *BS* zu bestimmen. Für die i-te Klasse $\mathscr{K}_i$ mit N_i Zellobjekten bedeutet dies:

$$BS_i = \frac{1}{N_i} \sum_{\{j \,|\, y_j \,\in\, \mathscr{K}_i\}} \left(y_j - \hat{y}_j^{(ens)} \right)^2 \,. \tag{6.7}$$

Dieser Score findet in den folgenden Modellstudien zur probabilistischen Bewertung der Güte der Ensemblevorhersagen von Zellobjekten einer bestimmten Klasse Verwendung.
Wird der Testdatensatz für jedes Ensemblemitglied variiert (vgl. Kapitel 6.1.1), bestimmt sich die Schwankungsbreite der Ensemblevorhersagen $\hat{\sigma}_j^{(ens)}$ in Gleichung (3.81) für das j-te Zellobjekt nur über die N'_{ens} Ensemblemitglieder, bei denen dieses im Testdatensatz vorhanden ist. Teilt man beispielsweise den Datensatz für jedes Mitglied zu etwa zwei Dritteln in Trainings- und zu einem Drittel in Testdaten auf, so wird jedes Zellobjekt im Mittel in etwa 0,66 N_{ens} Trainings- und 0,34 N_{ens} Testdatensätzen enthalten

sein (s. Kapitel 6.2.2). Unterschiedliche Zellobjekte treten allerdings in unterschiedlich vielen Testdatensätzen auf, wobei die Häufigkeit näherungsweise normalverteilt ist. Daher kann $\hat{\sigma}_j^{(ens)}$ mehr Werte annehmen als in dem Fall, in dem jedem Mitglied derselbe Testdatensatz zugrundeliegt (vgl. Kapitel 3.6.1).

6.2 Erste Modellstudie mit zwei Prädiktoren: DLS und LI

Die 38 553 Zellobjekte aus dem kombinierten Datensatz (Zellattribute und Umgebungsvariablen) werden nun in zwei Klassen aufgeteilt: solche mit kurzer und solche mit langer Lebensdauer. Auf den Klassentrennwert τ für die Lebensdauer (vgl. Kapitel 3.6.1, 5.3.1 und 6.1) und weitere Festlegungen für eine erste Modellstudie geht Kapitel 6.2.1 ein. Anschließend zeigt Kapitel 6.2.2 einen Vergleich der zwei verschiedenen nicht-linearen Klassifikationsverfahren: der logistischen Regression und des *Random Forests* (vgl. Kapitel 3.3.2 und 3.4). Dies geschieht in einem einfachen ersten Setup mit lediglich zwei Prädiktoren, der Modellstudie U2_0. Die Evaluation der Verfahren behandelt für diese Modellstudie einige Details, um mit der Interpretation der Ergebnisse vertraut zu werden und die vielen verschiedenen Aspekte hervorzuheben, über die anhand der Modellstudien relevante Informationen gewonnen werden können. Die Ergebnisse der daran anschließenden Modellstudien in den Kapiteln 6.3 und 6.4 werden dort in einer kompakten Weise vorgestellt und diskutiert.

6.2.1 Beschreibung der ersten Modellstudie

Um die angewendeten Verfahren detailliert zu analysieren und diskutieren, wird zunächst ein einfaches Setup mit lediglich zwei Umgebungsvariablen als Prädiktoren vorgestellt, die Modellstudie U2_0. Diese wird anschließend in den Kapiteln 6.3 und 6.4 in die Modellstudie U2 überführt und mit vielen

weiteren Modellstudien verglichen, die auf Kombinationen von bis zu 20 ausgewählten Prädiktoren basieren. Als Prädiktoren in der Modellstudie U2_0 fungieren zwei Umgebungsvariablen, die (a) relevant für die Entstehung bzw. Organisation hochreichender Konvektion sind, (b) möglichst unkorreliert sind (vgl. Kapitel 5.2.2), und (c) in den Analysen aus Kapitel 5.3 bereits unter den am besten zwischen Zellobjekten unterschiedlicher Lebensdauer unterscheidenden Variablen zu finden sind. So fällt die Wahl nach dem Testen verschiedener Kombinationen auf die DLS und den LI_{100hPa} als Ausdruck der vertikalen Windscherung bzw. der thermischen Instabilität der Atmosphäre. Beide Variablen haben sich bereits in verschiedenen Studien als geeignete Prädiktoren für hochreichende Konvektion gezeigt (z. B. Kunz, 2007; Púčik et al., 2015; Rädler et al., 2019; Kunz et al., 2020). Außerdem ist die (Rang-)Korrelation der beiden Variablen niedrig ($r_S = 0{,}15$ bzw. $r_P = 0{,}19$; vgl. Abbildung 5.13). Zudem gehört die DLS zu den Variablen, die in den Analysen in Kapitel 5.3.1 zum Unterscheidungsvermögen zwischen Zellobjekten unterschiedlicher Lebensdauer die besten kategorischen Gütemaße erreichen. Der LI_{100hPa} rangiert zwar etwas weiter hinten, dennoch ist die Wahl auch mit Blick auf die Abbildungen D.4c und 5.14 sinnvoll.

Die folgenden Festlegungen für die Modellstudie U2_0 wurden auf der Basis von einer großen Anzahl von Tests getroffen. Ausführliche Sensitivitätsuntersuchungen hierzu finden sich in Anhang B. Die Werte der beiden Prädiktoren, welche den Umgebungsvariablen zum Zeitpunkt der ersten Detektion des jeweiligen Zellobjekts entsprechen, werden einer z- und einer Yeo-Johnson-Transformation unterzogen (vgl. Kapitel 3.5.1 und 6.1.1). Der Klassentrennwert τ für die Unterscheidung in Zellobjekte mit kurzer oder langer Lebensdauer wird auf $\tau = 1\,\mathrm{h} = 60\,\mathrm{min}$ festgelegt. Damit zählen im Datensatz von U2_0 von den insgesamt 38 553 Zellobjekten 1 096 zur Klasse lange Lebensdauer (L) und 37 457 zur Klasse kurze Lebensdauer (K). Das Klassenverhältnis ρ_K, also die Anzahl von Zellobjekten mit einer langen Lebensdauer geteilt durch die Anzahl von Zellobjekten mit einer kurzen Lebensdauer, liegt daher bei etwa 2,9 % (vgl. Kapitel 3.5.2). Als Testdatensatz

dienen 34 % des kompletten Datensatzes. Um systematische Abhängigkeiten zu verhindern, die alleine auf die Datenauswahl zurückzuführen wären, wird dazu zufällig (ohne Zurücklegen) aus dem gesamten Datensatz der 38 553 Zellobjekte gezogen, sodass die ursprüngliche Verteilung der Lebensdauer und somit das Klassenverhältnis in sehr guter Näherung erhalten bleibt. Als halbe Breite des Übergangsbereichs für den Klassentrennwert τ bei der Evaluation dient $\tau' = 15$ min, sodass Zellobjekte mit einer Lebensdauer zwischen $\tau - \tau' = 45$ min und $\tau + \tau' = 75$ min nicht in die Berechnung der Gütemaße eingehen. Dies modifiziert das Klassenverhältnis im Testdatensatz im Vergleich zur Verteilung des gesamten Datensatzes leicht (s. u.). Das Ziehen der Trainingsdaten erfolgt durch ein *Bootstrapping* aus den verbleibenden 66 % der Daten ($f_{Tr} = 0{,}66$). Die Stichprobengröße des Trainingsdatensatzes ist $N_{Tr} = 25\,000$. Eine Balancierung des Trainingsdatensatzes bezüglich der Lebensdauer erfolgt vorerst nicht (vgl. Kapitel 3.5.2 und 6.1.1).
Mit den beiden Vorhersagemethoden (logistische Regression und *Random Forest*) werden zunächst jeweils zehn verschiedene Modelle trainiert, die auf unterschiedlichen Trainingsdaten basieren. In den folgenden Abbildungen und Tabellen werden die Methoden aufgrund des häufigen Auftretens mit LOGR und RF abgekürzt. Der Algorithmus für das *Bootstrapping* wählt für das Training der Modelle unterschiedliche Startwerte für die Pseudo-Randomisierung beim Ziehen. Je nach Untersuchung setzen sich die Testdaten für alle Modelle entweder aus denselben Zellobjekten oder einer unterschiedlichen Auswahl von Zellobjekten zusammen. Dies wird im Folgenden jeweils im Detail spezifiziert. Aufgrund des niedrigen Klassenverhältnisses, d. h. des starken Ungleichgewichts der Repräsentation von Zellobjekten mit kurzer und langer Lebensdauer im Datensatz, ist eine Variation der Entscheidungstrennwerte im niedrigen Bereich notwendig: $\mu_{LOGR} \in [0{,}01\,; 0{,}06]$ und $\mu_{RF} \in [0{,}001\,; 0{,}101]$. Für jedes Modell werden 51 Realisierungen mit unterschiedlichen Entscheidungstrennwerten berechnet. Der *Random Forest*

besteht aus $N_{Baum} = 1\,000$ Bäumen. Da es lediglich zwei Prädiktoren gibt, für welche die Minimierung der Residuen an einem Splitpunkt durchgeführt werden kann, ist $N_{split} = 2$ eine sinnvolle Wahl (vgl. Kapitel 3.4.3).

6.2.2 Evaluation der ersten Modellstudie

Das in Kapitel 3.6.1 beschriebene ROC-Diagramm hilft, einen Überblick über verschiedene Realisierungen der logistischen Regression und des *Random Forests* zu bekommen, die auf unterschiedlichen Entscheidungstrennwerten μ bezüglich $\hat{p}(y = 1 \mid \mathbf{x} = \mathbf{x}_j)$ bei der Zuordnung der Vorhersagen zu den beiden Klassen des Prädiktanden beruhen (vgl. Gleichungen (3.36), (3.71) und (3.72)). Je nach Fragestellung können Vorhersagen von unterschiedlichen Realisierungen als beste Vorhersagen bewertet werden. Deswegen werden unter anderem folgende zentrale Fragestellungen in den Auswertungen diskutiert:

(A) Sollen die Vorhersageverfahren eher möglichst viele Zellobjekte mit langer Lebensdauer als solche erkennen? Dann empfiehlt sich eine Realisierung weiter rechts auf der ROC-Kurve. Dafür müssen womöglich deutlich mehr Fehlalarme in Kauf genommen werden.

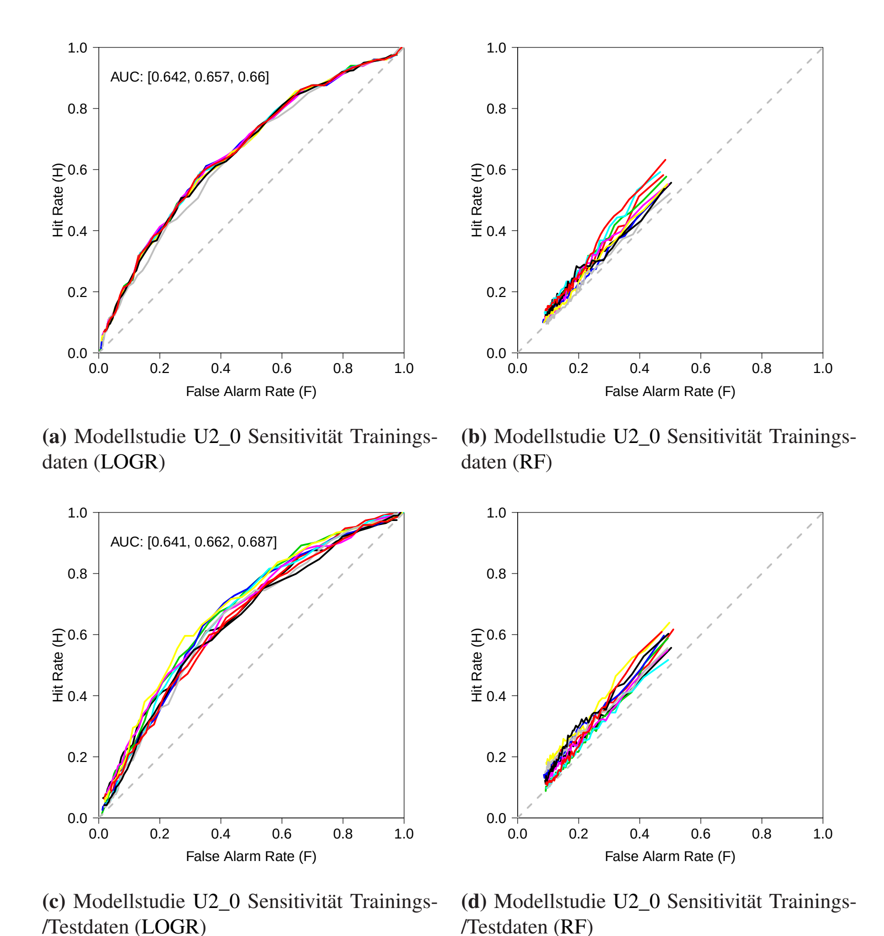
(B) Sollen die Vorhersageverfahren sowohl Zellobjekte mit langer als auch kurzer Lebensdauer möglichst gut vorhersagen? Dann empfiehlt sich eine Realisierung, die die TSS bzw. den Abstand zur Diagonalen im ROC-Diagramm D maximiert (vgl. Abbildung 3.4). Insbesondere für diese Fragestellung ist zudem die Betrachtung von balancierten Gütemaßen nützlich (Kapitel 6.1.2).

(C) Oder sollen die Vorhersageverfahren insgesamt möglichst viele Zellobjekte korrekt vorhersagen? Dann empfiehlt sich die Betrachtung weiterer Scores wie z. B. des *PC* oder des *False Alarm Ratios* (*FAR*), da H und F nicht sensitiv bezüglich des Klassenverhältnisses ρ_K sind (vgl. Tabelle 3.2).

Damit behandeln diese Fragestellungen mehrere Aspekte der Vorhersagen, die bei einer Entscheidung für einen der gezeigten Modellansätze als Anwendung in einem *Nowcasting*-Verfahren relevant sein könnten.

Deterministische Evaluation

Die ROC-Kurven aus den 51 Realisierungen für die jeweiligen Modelle zeigen, dass die Vorhersagen beider Verfahren als mäßig bis schlecht einzuordnen sind (Abbildung 6.2). Die logistische Regression liefert allerdings bessere Vorhersagen als der *Random Forest*. Im Sinne der kategorischen (binären) Evaluation entspricht die Klasse L einem Ereignis und K einem Nicht-Ereignis (Kapitel 3.6.1; vgl. Tabelle 6.1). H stellt folglich den Anteil der Zellobjekte aus Klasse L dar, die das jeweilige Modell als zur Klasse L gehörig vorhersagt. F entspricht umgekehrt dem Anteil aus Klasse K, die das jeweilige Modell als zur Klasse L gehörig vorhersagt.
Für die Fläche unter der ROC-Kurve (AUC; vgl. Kapitel 3.6.1), die Werte zwischen 0 und 1 annehmen kann mit 1 als optimalem Wert, liefert die logistische Regression Werte von bis zu $AUC = 0{,}687$ (Abbildung 6.2c). Die Variabilität der Kurven der zehn verschiedenen Modelle bei der Variation des Trainingsdatensatzes ist im Fall der logistischen Regression nur gering (Abbildung 6.2a). Mehr Variabilität wird erfasst, wenn jedem Modell sowohl unterschiedliche Trainings- als auch Testdaten zugrunde liegen (Abbildung 6.2c). Aus diesem Grund werden in den später gezeigten Ensemblevorhersagen grundsätzlich beide Datensätze variiert. Ein *Overfitting* eines der Modelle, also eine Überanpassung an die verwendeten Trainingsdaten, ist hier nicht zu erkennen, da die ROC-Kurven sehr ähnlich verlaufen und stets links der Diagonalen liegen, die eine binäre zufällige Vorhersage charakterisiert. Die logistische Regression mit der DLS als einzigem Prädiktor führt bereits zu Werten von $AUC = 0{,}64$ und hat somit deutlich mehr Einfluss als der $\mathrm{LI}_{100\mathrm{hPa}}$, der alleine lediglich $AUC = 0{,}56$ aufweist. Dieses Ergebnis passt gut zu dem bereits analysierten Unterscheidungsvermögen der Variablen (Kapitel 5.3.1).



(a) Modellstudie U2_0 Sensitivität Trainingsdaten (LOGR)

(b) Modellstudie U2_0 Sensitivität Trainingsdaten (RF)

(c) Modellstudie U2_0 Sensitivität Trainings-/Testdaten (LOGR)

(d) Modellstudie U2_0 Sensitivität Trainings-/Testdaten (RF)

Abbildung 6.2: ROC-Kurven basierend auf 51 Realisierungen mit unterschiedlichen Entscheidungstrennwerten für zehn verschiedene Modelle (a,c) der logistischen Regression und (b,d) des *Random Forests*. Die Modelle sind farblich codiert. (a) und (b) stellen eine Modellstudie mit variierenden Trainingsdaten mit jeweils denselben Testdaten dar. (c) und (d) stellen eine Modellstudie mit variierenden Trainings- und Testdaten dar. Zusätzlich ist in (a) und (c) die minimale, mittlere und maximale *AUC* angegeben.

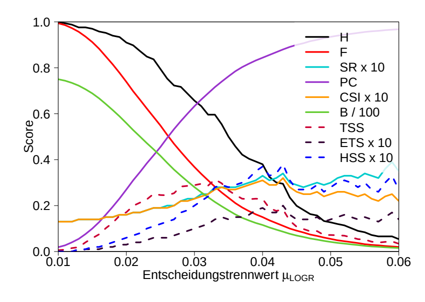
Tabelle 6.1: Kontingenztabellen für die Vorhersage der Lebensdauer der Zellobjekte als Klassifikation Kurz/Lang ($\tau = 60\,\text{min}$) mittels der logistischen Regression (oben; $\mu_{LOGR} = 0{,}029$) und des *Random Forests* (unten; $\mu_{RF} = 0{,}002$) für Modell G (s. u.; Modellstudie U2_0). Da $\tau' = 15\,\text{min}$ gewählt wurde, ist hier die Anzahl der evaluierten Zellobjekte $N'_{Te} = 12\,526 < 13\,108 = (1 - f_{Tr})N$ (vgl. Kapitel 3.6.1).

Beobachtung → Vorhersage ↓	Lang (L)	Kurz (K)
Lang (L)	$a = 114$	$b = 4\,950$
Kurz (K)	$c = 52$	$d = 7\,410$

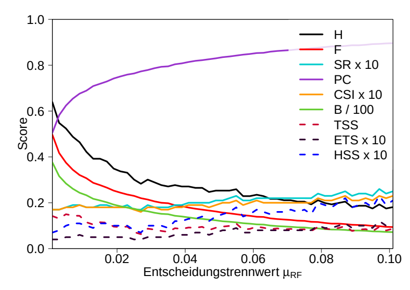
Beobachtung → Vorhersage ↓	Lang (L)	Kurz (K)
Lang (L)	$a = 91$	$b = 5\,155$
Kurz (K)	$c = 75$	$d = 7\,205$

Da der *Random Forest* seine Stärken bei Vorhersagen mit einer großen Anzahl von Prädiktoren hat (vgl. Kapitel 6.1.1), sind erwartungsgemäß die Vorhersagen mit lediglich zwei Prädiktoren kaum besser als eine binäre zufällige Vorhersage. Die einzelnen Realisierungen der zehn Modelle streuen unregelmäßiger als die der logistischen Regression, sodass eine sinnvolle Berechnung der *AUC* nicht möglich ist (Abbildungen 6.2b und 6.2d). Zudem zeigt sich eine etwas höhere Variabilität der Kurven der zehn verschiedenen *Random Forest*-Modelle. Man beachte, dass $\mu_{RF} = 0{,}001$ bei 1 000 Bäumen der kleinstmögliche Entscheidungstrennwert ist. Eine Vergrößerung der Baumanzahl auf z. B. 3 000 mit noch geringeren Trennwerten ändert jedoch qualitativ nichts an den gezeigten Ergebnissen (nicht gezeigt).

Für ein ausgewähltes Modell (als Modell G bezeichnet, gelbe Linie in Abbildung 6.2c bzw. 6.2d) findet sich in Abbildung 6.3 ein Überblick über verschiedene Gütemaße der kategorischen Evaluation, die sich für jede der 51 Realisierungen mit Hilfe der Kontingenztabelle (Tabelle 3.1) berechnen lassen. In Tabelle 6.1 sind für eine beispielhafte Realisierung von Modell G die Kontingenztabellen für die logistische Regression und den *Random Forest* gezeigt ($\mu_{LOGR} = 0{,}029$, $\mu_{RF} = 0{,}002$). Für die evaluierten Zellobjekte findet man ein starkes Ungleichgewicht von $\rho'_{K,Te} = (a+c)(b+d)^{-1} \approx 1{,}3\,\%$. Dies liegt noch unter dem Klassenverhältnis des gesamten Datensatzes von $\rho_K \approx 2{,}9\,\%$, da aufgrund der Wahl von $\tau' = 15\,\text{min}$ die Zellobjekte nahe des Klassentrennwerts τ nicht in die Evaluation eingehen (vgl. Kapitel 6.2.1).
Es ist daher nicht verwunderlich, dass das Fehlalarmverhältnis mit Werten jenseits von $FAR = 0{,}93$ (bezogen auf alle Realisierungen der zehn Modelle) generell sehr hoch und das *Success Ratio* (SR) mit $SR < 0{,}07$ sehr niedrig ist, da L-Vorhersagen für die Zellobjekte mit kurzer Lebensdauer beide Gütemaße dominieren. Zum Vergleich: Wird immer eine lange Lebensdauer vorhersagt, ist $FAR = 1 - \rho'_{K,Te}$; für $\tau = 60\,\text{min}$ ist folglich $FAR \approx 0{,}987$. Dies manifestiert sich auch in hohen Werten für den Bias (B), die für akzeptable Werte von H ein starkes *Overforecasting* anzeigen. Der *Proportion Correct* (PC) wird mit steigendem Entscheidungstrennwert größer, da für mehr Zellobjekte mit kurzer Lebensdauer auch K-Vorhersagen getroffen werden, allerdings auf Kosten von H. Der Heidke Skill Score (HSS), der *Equitable Threat Score* (ETS) und der *Critical Success Index* (CSI) verhalten sich qualitativ ähnlich und deuten die höchste Güte für hohe Entscheidungstrennwerte an, d. h. sie bewerten die Reduzierung des FAR höher als eine gute Vorhersage der Zellobjekte mit langer Lebensdauer, an der man jedoch im Hinblick auf die gefährlichen Begleiterscheinungen konvektiver Zellen eher interessiert ist (s. o. Fragestellung (A); vgl. Kapitel 2.2). Im Fall der logistischen Regression findet sich das Maximum der *True Skill Statistic* (TSS) im Mittel bei der zwanzigsten der 51 Realisierungen ($\mu_{LOGR} = 0{,}029$), bei der das *Odds*



(a) Modellstudie U2_0 Sensitivität μ (Modell G; LOGR)



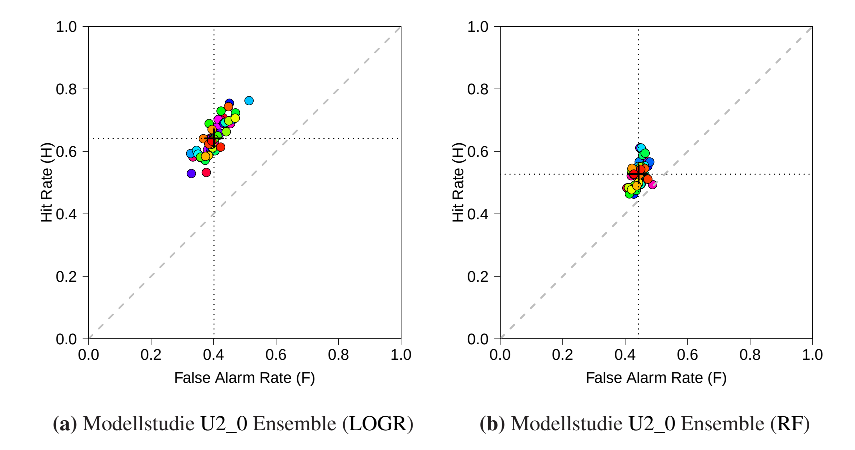
(b) Modellstudie U2_0 Sensitivität μ (Modell G; RF)

Abbildung 6.3: Verschiedene (Skill) Scores basierend auf 51 Realisierungen mit unterschiedlichen Entscheidungstrennwerten μ für jeweils ein repräsentatives Modell (a) der logistischen Regression und (b) des *Random Forests*. Einige der gezeigten Gütemaße sind zur besseren Übersicht skaliert dargestellt.

Ratio (*OR*) bei etwa $OR \approx 3$ liegt (nicht gezeigt). Die Chance, Zellobjekte mit langer Lebensdauer nach L-Vorhersagen zu beobachten, ist dort demnach etwa dreimal höher als nach einer K-Vorhersage.

Probabilistische Evaluation – Ein Ensembleansatz

Für einen Entscheidungstrennwert μ, der H und $1-F$ in etwa ausbalanciert sowie einen hohen TSS-Wert erreicht, folgt im Anschluss eine Ensemblestudie mit 51 Modellen für U2_0. Dieses Ensemble orientiert sich damit an der eingangs dieses Abschnitts beschriebenen Fragestellung (B), bei der man daran interessiert ist, sowohl Zellobjekte mit langer als auch kurzer Lebensdauer möglichst gut vorherzusagen. Unterschiedliche Vorhersageverfahren erfordern in der Regel unterschiedliche Entscheidungstrennwerte μ, um Ensemblevorhersagen zu generieren, deren Mittelwertspunkte ($F^{(ens)}$, $H^{(ens)}$) in etwa auf dem gleichen Lot der ROC-Diagonalen liegen. Die passenden Entscheidungstrennwerte sind hier $\mu_{LOGR} = 0{,}029$ sowie $\mu_{RF} = 0{,}002$. Damit das Ensemble möglichst viel Variabilität abdeckt, wird, wie in Kapitel 6.1.1 erläutert, für jedes Modell ein anderer potentieller



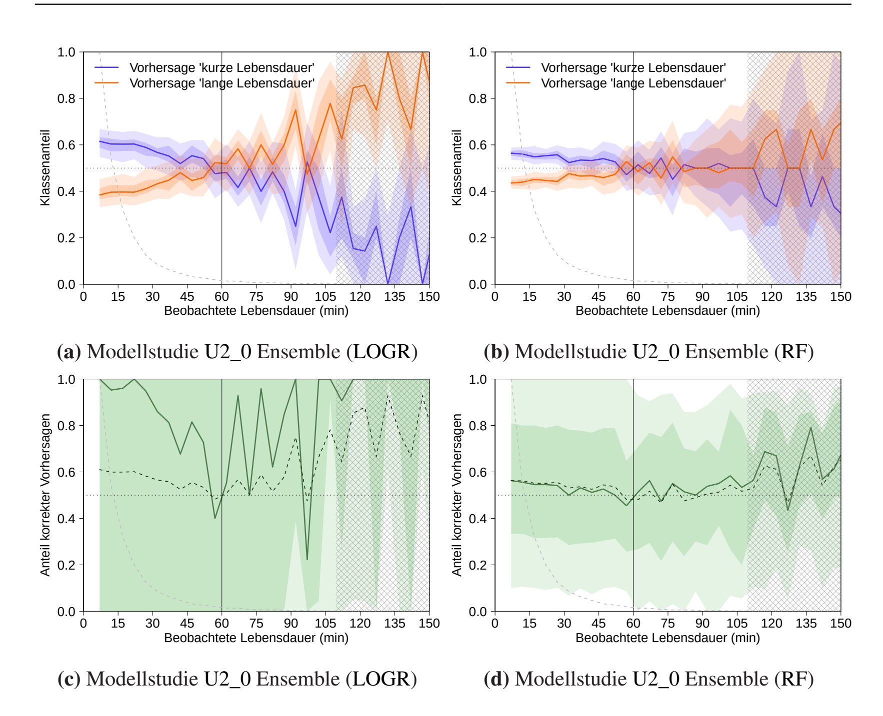
(a) Modellstudie U2_0 Ensemble (LOGR) **(b)** Modellstudie U2_0 Ensemble (RF)

Abbildung 6.4: Trefferrate (H) und Fehlalarmrate (F) für ein Ensemble aus 51 verschiedenen Modellen (a) der logistischen Regression und (b) des *Random Forests* bei festen Entscheidungstrennwerten $\mu_{LOGR} = 0{,}029$ bzw. $\mu_{RF} = 0{,}002$ bei gleichzeitiger Variation von Trainings- und Testdatensatz (bunte Kreise). Zusätzlich ist das Ensemblemittel mit einem schwarzen Kreuz und gestrichelten Linien dargestellt.

Trainings- und Testdatensatz verwendet. Der Ensembleansatz bietet zudem die Möglichkeit einer probabilistischen Vorhersage der Lebensdauer-Klasse einzelner Zellobjekte und liefert somit ein Maß für die Eintrittswahrscheinlichkeit von hohen Werten für die Lebensdauer. Damit lassen sich Werte für den *BS* berechnen (vgl. Kapitel 3.6.1). Darüber hinaus werden die klassenspezifischen *BS* bestimmt (vgl. Kapitel 6.1.2).

Bedingte Evaluation

H und F variieren innerhalb des Ensembles teils deutlich (Abbildung 6.4). Dies unterstreicht, dass der Ensembleansatz ein wichtiger Schritt für eine differenzierte Evaluation ist. Wie in Kapitel 6.1.2 erläutert, wird im Folgenden auch eine bedingte Evaluation, d. h. eine separate Auswertung für jede beobachtete Lebensdauer, vorgenommen. Die Vorhersageverfahren prognostizieren Zellobjekte mit einer Lebensdauer, die weit entfernt vom



(a) Modellstudie U2_0 Ensemble (LOGR) **(b)** Modellstudie U2_0 Ensemble (RF)

(c) Modellstudie U2_0 Ensemble (LOGR) **(d)** Modellstudie U2_0 Ensemble (RF)

Abbildung 6.5: Bedingte Quantil-Plots für die Anteile aller Zellobjekte gleicher beobachteter Lebensdauer an den beiden Vorhersage-Klassen für die Lebensdauer (K: violett, L: orange) für dasselbe Ensemble wie in Abbildung 6.4 mittels (a) der logistischen Regression und (b) des *Random Forests*. Sowie: Bedingte Quantil-Plots für die Anteile korrekter Prognosen für jedes einzelne Zellobjekt $\hat{y}_j^{(ens)}$ – zusammengefasst für Zellobjekte gleicher beobachteter Lebensdauer – für dasselbe Ensemble mittels (c) der logistischen Regression und (d) des *Random Forests*. Dargestellt sind Median (Linie), der Interquartilsbereich (dunkle Schattierung), das 5. und 95. Perzentil (helle Schattierung), sowie in (c) und (d) der Mittelwert, der gemäß Gleichung (6.2) durch *acc* gegeben ist (schwarz gestrichelte Linie). Gekreuzt sind die Bereiche, in denen weniger als 20 Zellobjekte dieselbe Lebensdauer erreichten. Die grau gestrichelte Linie veranschaulicht den jeweiligen Anteil der Zellobjekte gleicher Lebensdauer an der Gesamtzahl von Zellobjekten, skaliert bzgl. der Objektanzahl mit einer Lebensdauer von 7 min.

Klassentrennwert liegt, insgesamt besser (Abbildungen 6.5a und 6.5b). Aufgrund der deutlich geringeren Anzahl von Zellobjekten mit langer Lebensdauer im Testdatensatz ist die Varianz zwischen den einzelnen Ensemblemitgliedern dort größer. Betrachtet man die Anteile korrekter

Vorhersagen für jedes Zellobjekt separat (Abbildungen 6.5c und 6.5d), fällt sofort die große Schwankungsbreite zwischen den Zellobjekten auf, insbesondere bei der logistischen Regression. Es gibt folglich Zellobjekte einer bestimmten beobachteten Lebensdauer, die sehr viele oder sogar alle Ensemblemitglieder korrekt vorhersagen. Andererseits gibt es auch viele Zellobjekte, die kaum ein oder gar kein Ensemblemitglied richtig vorhersagt. Von den Zellobjekten mit kurzer Lebensdauer erreicht die Gruppe der Zellobjekte mit vergleichsweise sehr kurzer Lebensdauer über Gleichung (6.2) die höchsten Werte für die Genauigkeit *acc*, während von den Zellobjekten mit langer Lebensdauer die Gruppe der Zellobjekte mit sehr langer Lebensdauer die höchsten Werte verzeichnet.

Bezugnehmend auf die oben beschriebenen Fragestellungen (A) und (B) können drei ausgewählte Gütemaße als grafischer Vergleich zwischen den beiden Vorhersageverfahren dienen: 1) Der *BS* der Zellobjekte mit langer Lebensdauer, um Fragestellung (A) zu bearbeiten; 2) die *ACC* für Fragestellung (B); und 3) $\hat{\bar{\sigma}}_{ens}$ zur Quantifizierung der mittleren Schwankungsbreite des jeweiligen Ensembles (Abbildung 6.6). Ähnlich wie beim ROC-Diagramm gilt in dieser Darstellung: Je weiter links oben das Symbol liegt, desto besser ist die Vorhersage zu bewerten.

Brier Scores

Eine konstante Vorhersage, die für alle Zellobjekte immer eine kurze Lebensdauer prognostiziert, ergibt bei der Betrachtung des gesamten Datensatzes mit Abstand den niedrigsten *BS* ($BS = 0{,}03$), da mehr als 98 % der evaluierten Zellobjekte eine kurze Lebensdauer aufweisen (Tabelle E.2). Folglich dominieren die Zellobjekte mit kurzer Lebensdauer den *BS* des Ensemblemittels (Eintrittswahrscheinlichkeit einer langen Lebensdauer) der logistischen Regression. Die Untersuchung für die Zellobjekte mit langer Lebensdauer alleine ($BS_L = 0{,}17$) zeigt, dass die logistische Regression besser ist als eine 50 %-Vorhersage ($BSS_L = 0{,}32$) bzw. eine zufällige Vorhersage, bei der jedem

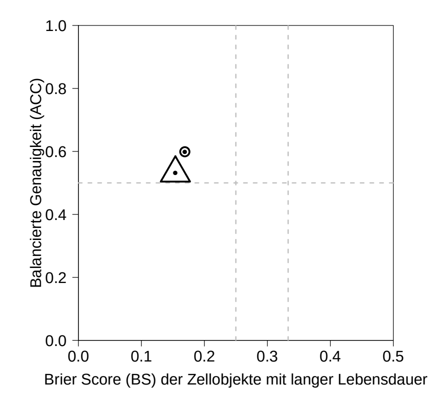


Abbildung 6.6: Synopse verschiedener Gütemaße für die Ensembles der logistischen Regression (Kreis) und des *Random Forests* (Dreieck; Modellstudie U2_0 Ensemble). Je größer die Symbole sind, desto größer ist die mittlere Schwankungsbreite des Ensembles $\hat{\bar{\sigma}}_{ens}$. Grau gestrichelte Linien stellen die Werte der Scores für zufällige Vorhersagen (horizontale + rechte vertikale Linie) bzw. einer 50 %-Vorhersage (linke vertikale Linie) dar.

Zellobjekt als Vorhersage zufällig ein Wert aus einer uniformen Verteilung der Eintrittswahrscheinlichkeiten zugewiesen wird ($BSS_L = 0{,}49$). Damit ist sie zugleich deutlich besser als eine zufällige Vorhersage, bei der man aus der beobachteten Verteilung von Zellobjekten mit kurzer und langer Lebensdauer zieht ($BSS_L = 0{,}83$). Aus dieser Perspektive kann man dem Ensemble der logistischen Regression folglich eine gewisse, wenn auch geringe probabilistische Vorhersagegüte attestieren.

Der BS bezüglich aller Zellobjekte ist mit $BS = 0{,}26$ für den *Random Forest* geringer im Vergleich zur logistischen Regression (Tabelle E.3). Die Erklärung dafür ist die geringere Schwankungsbreite $\hat{\bar{\sigma}}_{ens}$ der logistischen Regression sowie die quadratische Natur von BS. Die Ensemblevorhersagen der logistischen Regression fallen bei der Berechnung von BS daher stärker ins Gewicht. Somit erreicht der *Random Forest* auch höhere Werte für die

verschiedenen *BSS* als die logistische Regression, ebenso bei der separaten Betrachtung der beiden Klassen der Lebensdauer.

Balancierte Genauigkeit

Für die logistische Regression ergibt sich unter Berücksichtigung von $\tau' = 15\,\text{min}$ gemäß Gleichung (6.1) eine balancierte Genauigkeit von $ACC \approx 0{,}599$, wobei nur solche 5 min-Intervalle in die Berechnung eingehen, in denen mehr als 20 Zellobjekte vorzufinden sind (vgl. Abbildung 6.5c). Die *ACC* erreicht einen Wert von 0,637, wenn man die Forderung nach der Mindestanzahl von Zellobjekten auf beispielsweise zehn Zellobjekte abschwächt, da dann weitere Intervalle höherer Lebensdauer in die Berechnung der *ACC* eingehen. Für den *Random Forest* liegt die balancierte Genauigkeit für eine Mindestanzahl von 20 Zellobjekten mit $ACC \approx 0{,}531$ erwartungsgemäß unter derjenigen der logistischen Regression (vgl. Abbildung 6.5d). In weiteren Untersuchungen mit höheren Werten für μ_{RF} zeigt sich, dass die *ACC* nahezu konstant bleibt. Das Ensemblemittel des *Proportion Correct* $PC^{(ens)}$ hingegen steigt, da (viel) mehr Zellobjekte mit kurzer Lebensdauer richtig und nur eine geringe Anzahl von Zellobjekten mit langer Lebensdauer weniger falsch vorhergesagt werden (nicht gezeigt). Dies verdeutlicht, dass die *ACC* besser für einen Vergleich der Verfahren geeignet ist, um alle Werte für die Lebensdauer gleichermaßen zu bewerten.

Mittlere Schwankungsbreite des Ensembles

Eine genauere Überprüfung der mittleren Schwankungsbreite für die logistische Regression zeigt, dass die über die Zellobjekte gemittelte Schwankungsbreite der Ensemblevorhersagen gemäß Gleichung (3.80) lediglich bei $\hat{\overline{\sigma}}_{ens} \approx 0{,}070$ liegt. Die Vorhersagen der einzelnen Modelle für dasselbe Zellobjekt unterscheiden sich also selten und daher liegen entweder die meisten Mitglieder richtig oder die meisten falsch. Nur solche Zellobjekte gehen in das Mittel ein, für welche mindestens zehn der 51 Ensemblemitglieder Vorhersagen treffen, was für mehr als 99 % der Objekte der Fall ist.

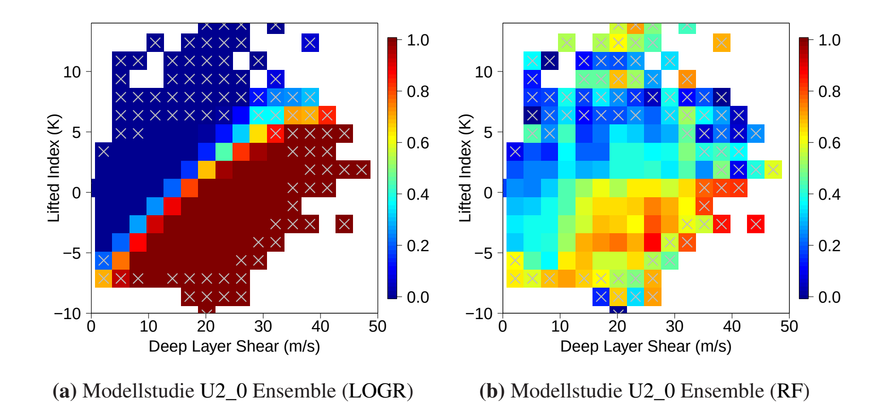
Aufgrund der zufälligen Variation von Trainings- und Testdatensatz ist die Verteilung der Häufigkeit der Zellobjekte näherungsweise normalverteilt mit $17{,}34 \pm 6{,}78$ Vorhersagen (Mittel über alle Objekte mit zweifacher Standardabweichung; vgl. Kapitel 6.1.2). Die Schwankungsbreite ist für den *Random Forest* mit $\hat{\bar{\sigma}}_{ens} \approx 0{,}387$ deutlich höher als bei der logistischen Regression, d. h. die unterschiedlichen *Random Forest*-Modelle sagen dasselbe Zellobjekt häufiger unterschiedlich vorher. Somit liegen die Anteile korrekter Vorhersagen der Ensemblemitglieder $\hat{y}_j^{(ens)}$ häufiger im mittleren Bereich zwischen 0 und 1. Dies zeigt sich beispielsweise in dem engeren Interquartilsbereich und der näheren Lage der Mittelwert-Linie (*acc*) an der Median-Linie in Abbildung 6.5d.

Zusammenfassend lässt sich konstatieren, dass die logistische Regression bessere Werte für deterministische Scores im Ensemblemittel (z. B. $H^{(ens)}$, $F^{(ens)}$) sowie eine höhere balancierte Genauigkeit *ACC* als der *Random Forest* erreicht. Letzterer wiederum reagiert sensibler auf den jeweiligen Trainings- und Testdatensatz, weswegen das Ensemble stärker in seinen Vorhersagen für einzelne Zellobjekte variiert. Dadurch ergeben sich bessere Werte für probabilistische Scores (z. B. *BS* bzw. *BSS*) mit dem *Random Forest* als mit der logistischen Regression.

Einfluss und Wichtigkeit der Prädiktoren

Bedingte Evaluation

In der gleichen Darstellungsweise wie in Abbildung 6.5 können die Vorhersagen nach den Werten der Prädiktoren analysiert werden (Abbildungen D.5 und D.6). So kann nachvollzogen werden, welche Werte der Prädiktoren zu welchen Vorhersagen führen und ob die statistischen Modelle dabei Abweichungen von dem meteorologisch erwartbaren Verhalten zeigen. Damit können die Modellvorhersagen auch physikalisch interpretiert werden und



(a) Modellstudie U2_0 Ensemble (LOGR) **(b)** Modellstudie U2_0 Ensemble (RF)

Abbildung 6.7: Mittlere Ensemblevorhersage für die Eintrittswahrscheinlichkeit einer langen Lebensdauer, aufgeteilt in verschiedene Gruppen der Prädiktoren DLS und LI_{100hPa}, für (a) die logistische Regression und (b) den *Random Forest*. Graue Kreuze geben Gruppen an, in denen 20 oder weniger Zellobjekte vorliegen.

das Vertrauen in die Modellvorhersagen gestärkt (oder geschwächt) werden. Qualitativ zeigt sich bei der logistischen Regression und dem *Random Forest* in der Modellstudie U2_0 dasselbe erwartbare Verhalten: Der Anteil von Zellobjekten, die eine L-Vorhersage erhalten, wächst mit steigender vertikaler Windscherung und zunehmender Instabilität (sinkendem LI_{100hPa}; vgl. Abbildungen D.5a+b und D.6a+b). Interessanterweise erreicht die Vorhersage der *Random Forests* ab etwa $DLS = 12\,m\,s^{-1}$ eine Art Plateau, sodass beide Klassen ähnlich häufig vorhergesagt werden. Bei niedriger Windscherung ist die Vorhersage demnach eindeutiger (meist eine K-Vorhersage). In Kombination mit dem *Lifted Index* wird jedoch deutlich, dass bei mittlerer bis hoher Scherung positive (negative) Werte des LI_{100hPa} deutlich niedrigere (höhere) Eintrittswahrscheinlichkeiten für eine lange Lebensdauer hervorrufen (Abbildung 6.7b).

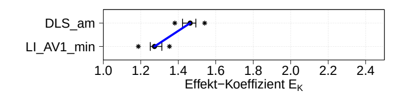
Der Anteil von Zellobjekten, die viele Ensemblemitglieder korrekt vorhersagen, sinkt mit steigender Scherung und steigender Instabilität (vgl. Abbildungen D.5c+d, D.6c+d). Dies liegt an dem weiterhin hohen Anteil von Zellobjekten mit kurzer Lebensdauer auch bei Umgebungsbedingungen, welche die Entwicklung organisierter Konvektion begünstigen (vgl. Kapitel 5.3.2). Viele Zellobjekte mit kurzer Lebensdauer erhalten dadurch fälschlicherweise eine L-Vorhersage. Die deutlich höhere Schwankungsbreite des *Random Forest*-Ensembles ist auch hier wieder erkennbar. Die Vorhersagen des Ensembles der logistischen Regression hingegen spiegeln den sigmoidalen Charakter der Wahrscheinlichkeitsfunktion wider, welcher auch in der kombinierten Darstellung der Eintrittswahrscheinlichkeiten für eine lange Lebensdauer gut zu erkennen ist (Abbildung 6.7a).
Mit dieser Analyse erhält man darüber hinaus einen Einblick in das deutliche *Overforecasting* von Zellobjekten mit langer Lebensdauer (s. o.): Die tatsächlich beobachteten Eintrittswahrscheinlichkeiten hierfür liegen für über 95 % aller in Abbildung 6.7 gezeigten Gruppen bei unter 5 %. Die Ensembles des *Random Forests* und der logistischen Regression geben jedoch für viele Gruppen deutlich höhere Eintrittswahrscheinlichkeiten vor – wenn auch auf recht unterschiedliche Art und Weise. Grund hierfür ist die Wahl der jeweiligen Entscheidungstrennwerte $\mu_{LOGR} = 0{,}029$ und $\mu_{RF} = 0{,}002$, welche für das Ensemble a priori so gewählt wurden, dass Zellobjekte mit langer und kurzer Lebensdauer in etwa gleich gute Vorhersagen erhalten (s. o. Fragestellung (B)). Für größere Entscheidungstrennwerte verringern sich konsequenterweise die mittleren Vorhersagen für die Eintrittswahrscheinlichkeit einer langen Lebensdauer.

Wichtigkeit der Prädiktoren

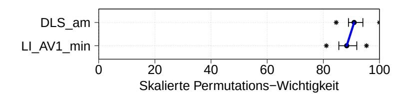
Als Nächstes erfolgt die Untersuchung der Wichtigkeit der Prädiktoren (*Predictor Importance*) für die Modelle des jeweiligen Vorhersageverfahrens. Diese gibt Aufschluss darüber, wie groß der relative Einfluss der Prädiktoren auf die Vorhersage in der jeweiligen Kombination ist. Dadurch wird

das Erstellen einer Rangliste der Prädiktoren ermöglicht, welche wiederum physikalisch interpretiert werden kann. Für die logistische Regression sind die Effekt-Koeffizienten E_K relevant, welche aufgrund der Anwendung der z-Transformation der Prädiktorwerte den standardisierten Effekt-Koeffizienten entsprechen (vgl. Kapitel 3.3.2). Für den *Random Forest* fällt hier die Wahl auf die Permutations-Wichtigkeit (vgl. Kapitel 3.4.3). Im Vergleich zur Gini-Wichtigkeit bringen beide Maße sehr ähnliche Reihenfolgen der Prädiktoren hervor, solange die Prädiktoren vom gleichen Typ sind (beispielsweise nur kontinuierliche Prädiktoren; Strobl et al., 2007). Die Permutations-Wichtigkeit ist meist etwas homogener verteilt als die Gini-Wichtigkeit (Hastie et al., 2009). Zur besseren Vergleichbarkeit mit weiteren Ensembles des *Random Forests* werden die Werte der Permutations-Wichtigkeit jeweils auf den von allen Prädiktoren angenommenen Maximalwert normiert und mit 100 multipliziert. Aufgrund der unterschiedlichen Methodik zur Bestimmung der Wichtigkeit der Prädiktoren ist ein Vergleich der Werte der Effekt-Koeffizienten mit denen der Permutations-Wichtigkeit nicht möglich. Vielmehr ist die jeweils relative Wichtigkeit der Prädiktoren innerhalb eines Vorhersageverfahrens sowie ein Vergleich der Reihenfolge der Prädiktoren zwischen den Vorhersageverfahren interessant. So kann festgestellt werden, welche Prädiktoren einen vergleichsweise großen Einfluss haben und ob sich systematische Gemeinsamkeiten in den unterschiedlichen Verfahren erkennen lassen, welche das Vertrauen in die Reihenfolge der Prädiktoren in den Ranglisten stärken.

In den bisher diskutierten Modellensembles beider Verfahren ist die vertikale Windscherung bezogen auf den Ensemblemedian die wichtigere Variable (Abbildung 6.8). Es gilt sogar $E_{K,DLS} > E_{K,LI}$ für jedes einzelne Modell der logistischen Regression, sodass sich die Wertebereiche der Effekt-Koeffizienten nicht überlappen. Im Gegensatz dazu hat bei etwa 18 % der *Random Forest*-Modelle die Instabilität eine höhere Wichtigkeit als die Windscherung und der Median der DLS liegt im Interquartilsbereich des $\mathrm{LI}_{100\mathrm{hPa}}$. Die Wichtigkeit der Prädiktoren stimmt folglich gut mit den Ergebnissen in



(a) Modellstudie U2_0 Ensemble (LOGR)



(b) Modellstudie U2_0 Ensemble (RF)

Abbildung 6.8: Wichtigkeit der Prädiktoren in den Ensembles (a) der logistischen Regression und (b) des *Random Forests*. Dargestellt sind der Median des Ensembles (schwarzer Punkt), der Interquartilsbereich (Balken) sowie das 5. und 95. Perzentil (Sterne), jeweils bezogen auf das entsprechende Maß (Effekt-Koeffizient bzw. skalierte Permutationswichtigkeit). Zusätzlich sind die Werte des Medians mit einer blauen Linie verbunden.

Kapitel 5.3.1 überein, in denen besonders dynamische Größen das größte Unterscheidungsvermögen zwischen Zellobjekten unterschiedlicher Lebensdauer zeigten.

6.3 Modellstudien zur Vorhersage der Lebensdauer

Dieses Kapitel stellt die Ergebnisse von Modellstudien vor, welche Vorhersagen für die Lebensdauer der Zellobjekte mittels Klassifikations- oder Regressionsmethoden für unterschiedliche Kombinationen von Prädiktoren treffen. Die ausführlichen Sensitivitätsuntersuchungen aus Anhang B geben Hinweise für die Datenvorbehandlung und geeignete Setups. Als Prädiktoren fungieren neben den Umgebungsbedingungen auch Zellattribute zu bestimmten Zeitpunkten zu Beginn der Zellentwicklung. Jede Modellstudie erhält zur Kennzeichnung eine Abkürzung, welche die Prädiktoren charakterisiert und im Akronymverzeichnis beschrieben ist. In den Unterkapiteln werden die Analysen der Modellstudien zunächst ausführlicher vorgestellt. Im Anschluss findet sich jeweils eine kurze Zusammenfassung und Interpretation der bedeutendsten Ergebnisse.

6.3.1 Evaluation von Klassifikationsverfahren zur Vorhersage der Lebensdauer

Die im Folgenden diskutierten Modellstudien für die logistische Regression und den *Random Forest* unterscheiden sich in der Anzahl und Auswahl der Prädiktoren. Das auf der ersten Modellstudie U2_0 basierende allgemeine Setup der Modelle der logistischen Regression und des *Random Forests* unterscheidet sich hauptsächlich darin, dass erstere generell kein *Resampling* erfahren, während für manche Studien mit dem *Random Forest* ein *Resampling* erfolgt. Letzteres wird angewendet, wenn die Ensemblemitglieder selbst mit einer großen Anzahl von Entscheidungsbäumen die Vorhersagen nicht scharf genug abbilden können (z. B. Abbildung B.3b für hohe Werte des Klassentrennwerts τ) und/oder das *Resampling* die Vorhersagen im Vergleich zu den Modellen ohne *Resampling* verbessert (s. u.). Die Wahl für die Entscheidungstrennwerte μ_{LOGR} bzw. μ_{RF} richtet sich in den Ensemblestudien wie im Ensemble von U2_0 nach der Balancierung von H und $1-F$. Nach einer Balancierung des Trainingsdatensatzes mittels eines *Resamplings* ist μ daher deutlich höher zu wählen als ohne Balancierung. Damit adressieren die Auswertungen insbesondere die Fragestellung (B) aus Kapitel 6.2.2 (Zellobjekte mit kurzer und langer Lebensdauer sollen gleich gut vorhergesagt werden). Als Ergänzung schließt sich am Ende dieses Abschnitts eine Studie an, welche die dortigen Fragestellungen (A) und (C) näher betrachtet (möglichst gute Vorhersagen von Zellobjekten mit langer Lebensdauer bzw. möglichst viele korrekte Vorhersagen).

Im Gegensatz zu U2_0 findet aufgrund der Sensitivitätsuntersuchungen kein *Bootstrapping* der Trainingsdaten statt (vgl. Anhang B). Das *Resampling* erfolgt als Kombination von *Undersampling* und *Oversampling* mit $\phi_{USP} = 0{,}65$, $N_{I,min} = 15$, $N_{USP} = N_{OSP} = 20$ und dem Gauss'schen Rauschen als *Oversampling*-Methode (vgl. Kapitel 3.5.1, 6.1.1 und Anhang B). Die gezeigten Modellstudien mit einem Ensemble des *Random Forests* basieren auf $N_{Baum} = 125$ Entscheidungsbäumen, während einige der nicht

gezeigten Voruntersuchungen ohne *Resampling* teilweise bis zu 2 000 Bäume verwendeten (s. o.; vgl. Kapitel 3.5.2). N_{split} entspricht jeweils den in Kapitel 3.4.3 und Anhang B genannten Standardwerten zur Klassifikation.

Im Folgenden werden drei verschiedene Kombinationsmöglichkeiten von Prädiktoren betrachtet (Tabelle 6.2):

(1) Kombinationen von zwei, sechs und 15 Umgebungsvariablen (U2, U6, U15).

(2) Kombinationen von zwei oder vier Zellattributen zu unterschiedlichen Zeitpunkten der Zellentwicklung (Z5, Z15, Z15$^+$).

(3) Kombinationen von zwei bzw. 15 Umgebungsvariablen und zwei bzw. vier Zellattributen (K5, K15, K15$^+$).

Um die Verwendung der Verlagerungsrichtung der Zellobjekte (zyklische Variable) als Prädiktor zu ermöglichen, ist die Bestimmung der horizontalen Komponenten $c_{Z,x}$ und $c_{Z,y}$ des Verlagerungsvektors $\mathbf{c}_Z$ zum jeweiligen Zeitpunkt notwendig. Zusammengefasst werden die Komponenten im Folgenden als Verlagerung der Objekte bezeichnet. Die Modellstudien Z5 und Z15 greifen nur auf die Information über die Zellfläche $A_Z(t = 7\,\text{min})$ bzw. $A_Z(t = 17\,\text{min})$ sowie die Fläche des Zellkerns $A_{Z,K}(t = 7\,\text{min})$ bzw. $A_{Z,K}(t = 17\,\text{min})$ zurück. Die Modellstudie Z15$^+$ hingegen berücksichtigt neben der Zell- und Zellkernfläche zusätzlich die Verlagerung der Zellobjekte. Der Verlagerungsvektor zum Zeitpunkt der zweiten Detektion durch KONRAD kann aufgrund der Berechnungsmethodik (vgl. Kapitel 4.3.2) nicht adäquat wiedergegeben werden. In Z15 und Z15$^+$ finden nur Zellobjekte Verwendung, die insgesamt mindestens viermal detektiert wurden (vgl. Kapitel 6.1.1).

Tabelle 6.2: Übersicht über die verschiedenen Kombinationen von Prädiktoren sowie die verwendeten Entscheidungstrennwerte μ für unterschiedliche Modellstudien zur Vorhersage der Lebensdauer. Oben: Modellstudien, die nur auf Umgebungsvariablen basieren; Mitte: Modellstudien, die nur auf Zellattributen basieren; unten: Modellstudien, die auf einer Kombination von Umgebungsvariablen und Zellattributen basieren. Alle Werte der Umgebungsvariablen entsprechen denjenigen zum Zeitpunkt der ersten Detektion der Zellen durch KONRAD.

Modellstudie →	U2	U6	U15
Prädiktoren	DLS, LI_{100hPa}	DLS, LI_{100hPa}, RH_{700hPa}, SRH_{0-3km}, NFK_{MU} T_{850hPa} (`sdam`; je 1 pro Cluster aus Abb. 5.14b)	15 beste Variablen aus Tabelle 5.1
μ_{LOGR}	0,029	0,029	0,029
μ_{RF}	0,640	0,600	0,580
Resampling für RF	✓	✓	✓

Modellstudie →	Z5	Z15	Z15^{+}
Prädiktoren	A_Z, $A_{Z,K}$ ($t = 7$ min)	A_Z, $A_{Z,K}$ ($t = 17$ min)	A_Z, $A_{Z,K}$, $c_{Z,x}$, $c_{Z,y}$ ($t = 17$ min)
μ_{LOGR}	0,029	0,075	0,075
μ_{RF}	0,550	0,640	0,050
Resampling für RF	✓	✓	×

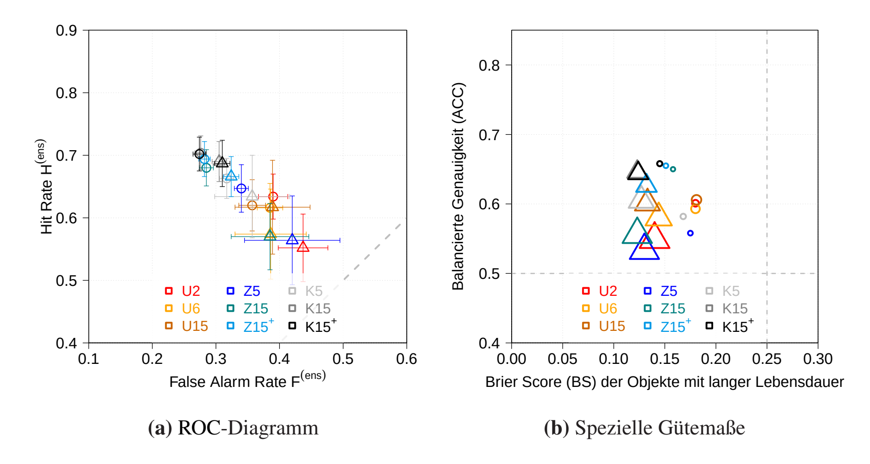
(Fortsetzung Tabelle 6.2)

Modellstudie →	K5	K15	$K15^{+}$	$K15^{+}_{var}$
Prädiktoren (LOGR)	wie U2 + Z5	wie U2 + Z15	wie U2 + $Z15^{+}$	wie $K15^{+}$
Prädiktoren (RF)	wie U15 + Z5	wie U15 + Z15	wie U15 + $Z15^{+}$	wie $K15^{+}$
μ_{LOGR}	0,029	0,075	0,075	[0,01 ; 0,21]
μ_{RF}	0,570	0,080	0,080	[0,01 ; 0,21]
Resampling für RF	✓	×	×	×

Detaillierte Analyse der verschiedenen Modellstudien

In Abbildung 6.9 sind Auswertungen von Ensemblemodellstudien mit den jeweils drei ausgewählten Kombinationen von Prädiktoren aus Tabelle 6.2 für die oben beschriebenen Kombinationsmöglichkeiten (1)–(3) dargestellt. Allgemein lassen sich zunächst folgende Auffälligkeiten bezüglich der Vorhersageverfahren festhalten (aufgrund des häufigen Auftretens finden die Abkürzungen LOGR und RF für die logistische Regression und den *Random Forest* in den nachfolgenden Beschreibungen Anwendung):

- Alle Vorhersagen erreichen bessere Scores als zufällige oder probabilistische Vorhersagen, die für jedes Zellobjekt eine Wahrscheinlichkeit von 50 % für eine kurze und 50 % für eine lange Lebensdauer ausgeben (grau gestrichelte Linien in Abbildung 6.9).
- Vorhersagen mit einer Kombination von Umgebungsvariablen und Zellattributen erreichen die besten Werte für die Gütemaße (s. u.).
- Die Vorhersagen der RF-Modellensembles weisen eine höhere Schwankungsbreite innerhalb des Ensembles $\hat{\overline{\sigma}}_{ens}$ auf als diejenigen der entsprechenden LR-Modellensembles (Größe der Symbole in Abbildung 6.9b; vgl. Kapitel 6.2.2).
- Die Schwankungsbreite hängt nur schwach von der Wahl der Prädiktoren ab.
- Die Schwankungsbreite sowie die Variabilität der Fehlerrate F (horizontale Balken in Abbildung 6.9a) ist für die Ensembles am größten, für die ein *Resampling* erfolgt, da dieses generell kleinere und je nach der genauen Verteilung der Werte der Lebensdauer unterschiedlich große reduzierte Trainingsdatensätze hervorbringt.
- Die RF-Vorhersagen erreichen eine höhere probabilistische Vorhersagegüte für die Zellobjekte mit einer langen Lebensdauer als die entsprechenden LOGR-Vorhersagen (Brier Score BS in Abbildung 6.9b; vgl. Kapitel 6.2.2).



(a) ROC-Diagramm

(b) Spezielle Gütemaße

Abbildung 6.9: Synopse von Modellstudien mit verschiedenen Kombinationen von Prädiktoren zur Vorhersage der Lebensdauer. Rötliche Farben markieren Studien, die nur Umgebungsvariablen als Prädiktoren verwenden und bläuliche solche, die nur Zellattribute verwenden. Graustufen kennzeichnen kombinierte Studien. Ensemblevorhersagen mittels eines Ensembles der logistischen Regression sind mit Kreisen, solche des *Random Forests* mit Dreiecken dargestellt. (a) Ensemblemittel und zugehöriger Variationsbereich für F und H im ROC-Diagramm – Balken entsprechen $\pm\sigma_{F,H}$; (b) analog zu Abbildung 6.6.

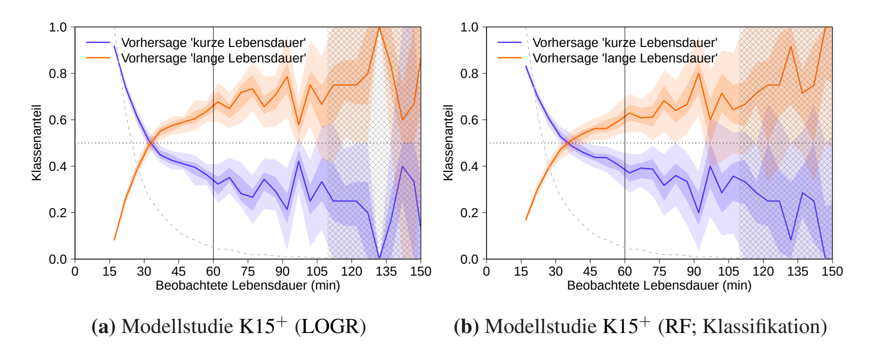
- Die Fehlalarmrate F variiert zwischen den Ensemblemitgliedern weniger stark als die Trefferrate H, wenn kein *Resampling* durchgeführt wird (Balken in Abbildung 6.9a).

Für die Modellstudien mit verschiedenen Kombinationen von Umgebungsvariablen als Prädiktoren (Tabelle 6.2; oben) lässt sich feststellen, dass sich die LOGR-Vorhersagen kaum verbessern, wenn mehr als zwei Umgebungsvariablen aus U2_0 verwendet werden (DLS und LI_{100hPa}). Die RF-Vorhersagen verbessern sich hingegen erwartungsgemäß mit zunehmender Anzahl von Umgebungsvariablen (vgl. Kapitel 6.1.1). Die Verwendung aller 33 statt 15 Umgebungsvariablen bringt dagegen nur eine minimale weitere Verbesserung der Gütemaße (nicht gezeigt). Die RF-Vorhersagen mit 15 Umgebungsvariablen erreichen im Ensemblemittel ähnlich gute Werte für die Gütemaße wie die LOGR-Vorhersagen (für den *BS* sogar bessere).

Für die Modellstudien mit verschiedenen Kombinationen von Zellattributen als Prädiktoren (Tabelle 6.2; Mitte) ist erkennbar, dass die LOGR- und RF-Vorhersagen mit der Information über die Zellfläche A_Z und die Fläche des Zellkerns $A_{Z,K}$ zum Zeitpunkt 15 min nach der ersten Detektion bessere Werte für die Gütemaße erreichen als solche mit der entsprechenden Information zum Zeitpunkt 5 min nach der ersten Detektion. Die Vorhersagen mit der zusätzlichen Information der Verlagerung des Zellobjekts verbessern die LOGR-Vorhersagen (RF-Vorhersagen) leicht (deutlich). Die LOGR-Vorhersagen erreichen allerdings im Ensemblemittel deutlich bessere Werte für die Gütemaße als die entsprechenden RF-Vorhersagen (außer für den *BS*).
Für die Modellstudien mit verschiedenen Kombinationen von Umgebungsvariablen und Zellattributen als Prädiktoren (Tabelle 6.2; unten) ergibt sich schließlich, dass die LOGR- und RF-Vorhersagen in fast allen Fällen bessere Werte für die Gütemaße erreichen als bei der jeweiligen Verwendung von Umgebungsvariablen und Zellattributen alleine (Ausnahme: Modellstudie K5 für LOGR; vgl. Abbildung 6.9b). Die LOGR-Vorhersagen erreichen im Ensemblemittel meist bessere Werte für die Gütemaße als die entsprechenden RF-Vorhersagen (wiederum außer für den *BS*). In $K15^+$ erreichen die LOGR- und RF-Vorhersagen allerdings eine ähnlich hohe balancierte Genauigkeit.

Zusammenfassende Analyse und Interpretation der Ergebnisse

Anhand dieser Analysen zusammen mit einer bedingten Evaluation für $K15^+$ (Abbildung 6.10), einer Synopse, bei der die Stärke der relativen Wichtigkeit der Prädiktoren zusammengefasst wird (Abbildung 6.11), sowie der ROC-Kurven für die Modellstudie $K15^+_{var}$ (Abbildung 6.12) lassen sich viele interessante Aspekte zur Verwendung der beiden Vorhersageverfahren zur groben Abschätzung der Lebensdauer (kurz/lang) ableiten. Aufgrund des großen Stichprobenumfangs können diese Ergebnisse als repräsentativ und robust für isolierte konvektive Zellen betrachtet werden.



(a) Modellstudie K15$^+$ (LOGR) **(b)** Modellstudie K15$^+$ (RF; Klassifikation)

Abbildung 6.10: Wie Abbildungen 6.5a+b, nur für K15$^+$.

- Ein Ensemble aus LOGR-Modellen (RF-Modellen) mit zwei (15) Umgebungsvariablen und vier Zellattributen ist in der Lage, im Mittel jeweils etwas über (unter) 70 % der evaluierten Zellobjekte korrekt einer der beiden Lebensdauerklassen zuzuordnen (K15$^+$; Abbildung 6.9a). Die balancierte Genauigkeit *ACC* erreicht für K15$^+$ nur Werte um 65 % (Abbildung 6.9b). Dies liegt insbesondere daran, dass beide Vorhersageverfahren bei einer Balancierung von H und $1-F$ Zellobjekte mit einer Lebensdauer von etwas unter 45 min bereits häufiger der Klasse von Objekten mit einer langen Lebensdauer zuordnen (Abbildung 6.10; vgl. Kapitel 6.1.2).
- Beide Vorhersageverfahren treffen mit der Information von nur vier Zellattributen zum Zeitpunkt 15 min nach der ersten Detektion als Prädiktoren bereits gute Vorhersagen (Z15$^+$). Die Hinzunahme von Umgebungsvariablen verbessert die Gütemaße zwar weiter, der Gewinn ist jedoch nicht allzu stark ausgeprägt (K15$^+$).

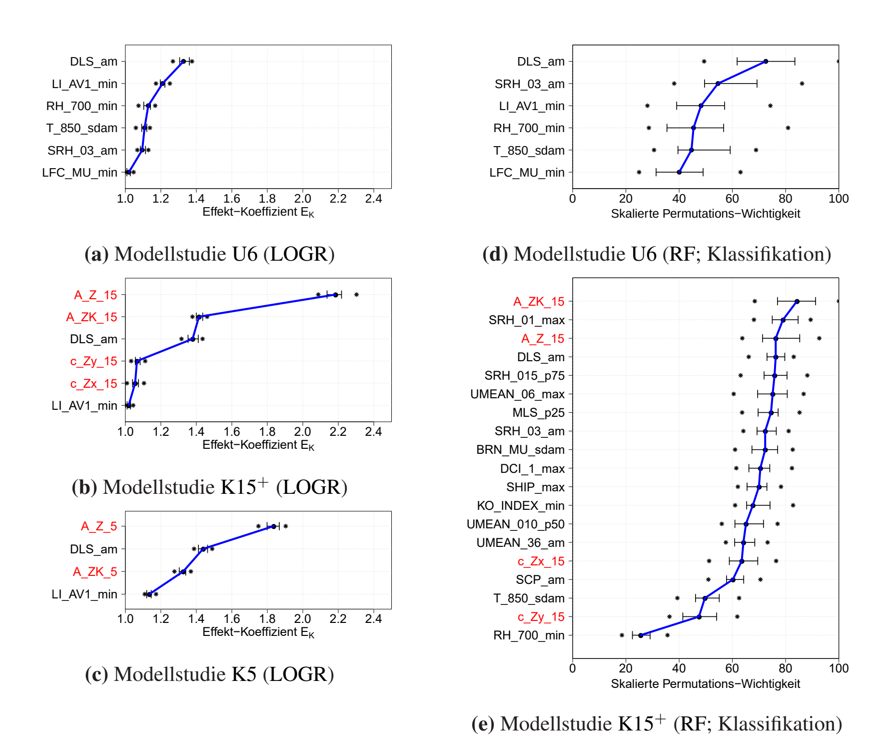


Abbildung 6.11: Wie Abbildung 6.8, nur für (a)+(d) U6, (b)+(e) K15$^+$ und (c) K5 (nur logistische Regression). Die Abkürzungen der Zellattribute A_Z, $A_{Z,K}$, $c_{Z,x}$ und $c_{Z,y}$ sind mit der Angabe der verstrichenen Zeit seit der ersten Detektion (min) in rot eingetragen, Umgebungsvariablen in schwarz.

- Die vertikale Windscherung, insbesondere die DLS, stellt die wichtigste Umgebungsvariable dar (Abbildung 6.11). In dem RF-Ensemble von K15$^+$ sind viele dynamische Variablen bedeutsamer im Vergleich zu thermodynamischen Variablen. Dies deckt sich mit den Analysen zum Unterscheidungsvermögen der Variablen aus Kapitel 5.3.1.
- Die Zellfläche A_Z ist für das LOGR-Ensemble das mit Abstand wichtigste Zellattribut sowie insgesamt der wichtigste Prädiktor (Abbildungen 6.11b+c). Die Fläche des Zellkerns $A_{Z,K}$ erreicht eine ähnliche

Wichtigkeit wie die Windscherung, was sich mit den Analysen zum Unterscheidungsvermögen der Variablen hinsichtlich der Lebensdauer deckt (vgl. Kapitel 5.3.1). Die Instabilität der Luftmasse und die Verlagerung der Zellobjekte spielen nur eine untergeordnete Rolle.

- Für das RF-Ensemble erreichen die Zellfläche und die Fläche des Zellkerns ähnlich hohe Werte für die Permutations-Wichtigkeit wie die Windscherung (Abbildung 6.11e). Wenig überraschend sind viele dynamische, stark miteinander korrelierte Variablen ähnlich wichtig. Auch hier ist die Verlagerung der Zellobjekte in Kombination mit den Umgebungsvariablen weniger relevant.
- Mehr als 86 % (80 %) der Zellobjekte mit einer langen Lebensdauer erhalten von einem LOGR-Ensemble (RF-Ensemble) eine korrekte Vorhersage, wenn man in Kauf nimmt, dass die Vorhersage der Zellobjekte mit einer kurzen Lebensdauer nicht besser als eine zufällige Vorhersage ist ($F = 50\,\%$; Abbildung 6.12; s. o. Fragestellung (A)).
- Eine simple Vorhersage, die immer eine kurze Lebensdauer vorhersagt, erreicht aufgrund des Ungleichgewichts des Datensatzes hinsichtlich der Lebensdauer die beste Vorhersagegüte, wenn diese sich auf den Anteil von korrekt vorhergesagten Zellobjekten bezieht (*PC*; s. o. Fragestellung (C)).

Die gezeigten Modellstudien stellen lediglich eine Auswahl der durchgeführten Untersuchungen dar. Viele der verwendeten Umgebungsvariablen sind durch andere, mit ihnen physikalisch verwandte und/oder statistisch stark korrelierte Variablen in den jeweiligen Studien ersetzbar. Externe Variablen wie beispielsweise die Tageszeit, Jahreszeit, der Längen- oder Breitengrad der registrierten Zellobjekte zeigen bei alleiniger Verwendung als Prädiktoren eine gewisse, wenn auch geringe Vorhersagegüte. Kombiniert man sie jedoch beispielsweise mit den Prädiktoren aus $K15^{+}$, lässt sich keine weitere Verbesserung der Vorhersagen der Ensembles der logistischen Regression und des *Random Forests* feststellen (nicht gezeigt).

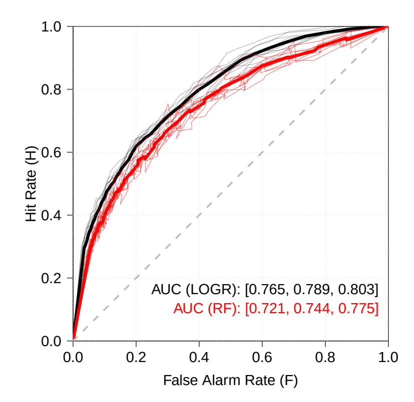


Abbildung 6.12: ROC-Kurven für zehn verschiedene Modelle der logistischen Regression (schwarz) und des *Random Forests* (rot) sowie deren mittlerer Verlauf (dicke Linie) zur Vorhersage der Lebensdauer in der Modellstudie $K15^{+}_{var}$.

6.3.2 Evaluation von Regressionsverfahren zur Vorhersage der Lebensdauer

Die im Folgenden diskutierten Modellstudien für zwei Regressionsverfahren, den Polynomansatz aus Kapitel 3.3.3 (Abkürzung in Abbildungen: POLY) und den *Random Forest*, basieren auf denselben Setups und derselben Auswahl von gezeigten Kombinationen der Prädiktoren wie die Klassifikationsverfahren in Kapitel 6.3.1[1]. Der entscheidende Unterschied ist, dass die beiden Regressionsverfahren nun eine kontinuierliche Vorhersage der Lebensdauer T_Z (Wert in Minuten) ermöglichen, statt nur eine Klassifizierung der Zellobjekte in zwei Klassen (kurze/lange Lebensdauer) vorzunehmen. Dabei findet dasselbe *Resampling* wie in der Datenvorbehandlung des *Random Forests* als Klassifikationsverfahren Anwendung. Die gezeigten Modellstudien

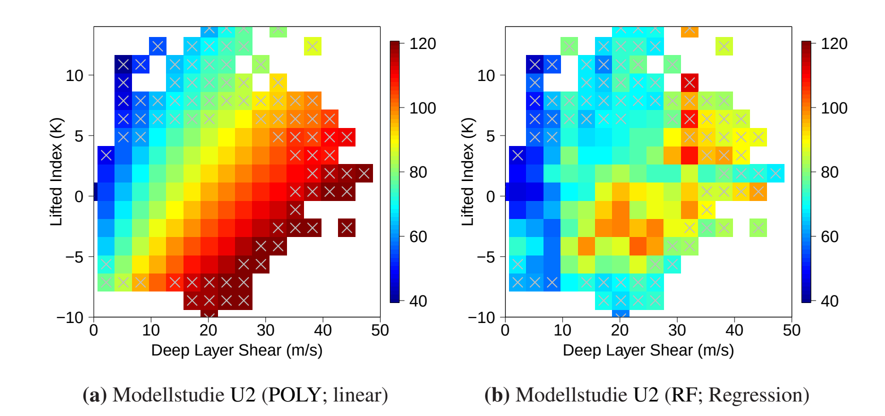
[1] Für den Polynomansatz werden in den kombinierten Modellstudien K5 und $K15^{+}$ stets dieselben Prädiktoren wie für den *Random Forest* gemäß Tabelle 6.2 ausgewählt.

mit einem Ensemble des *Random Forests* verwenden demnach dieselbe Anzahl von $N_{Baum} = 125$ Entscheidungsbäumen wie die Studien mit den Klassifikationsverfahren. N_{split} entspricht jeweils den in Kapitel 3.4.3 und Anhang B genannten Standardwerten für Regressionsverfahren.
In den Untersuchungen mit einem Ensemble des Polynomansatzes nehmen Instabilitäten der Vorhersagen mit zunehmender Anzahl an Prädiktoren zu, d. h. mit steigender Anzahl von unabhängigen Variablen und steigender Ordnung des Polynoms N_p (vgl. Kapitel 3.3.3). Umso größer sollte dementsprechend der Regularisierungsparameter α gewählt werden. In den folgenden Studien gilt überall einheitlich $\alpha = 0{,}01$, was die Vorhersagen mit Polynomen niedriger Ordnung kaum gegenüber den ungedämpften Lösungen verändert und solche höherer Ordnung adäquat dämpft, wie durch Untersuchungen mit verschiedenen Kombinationen von Prädiktoren und Werten von $\alpha \in [10^{-6}; 10^{0}]$ festgestellt wurde. Die Untersuchungen beinhalten im nicht-linearen Ansatz aus Gleichung (3.38) generell keine Mischterme von unterschiedlichen unabhängigen Variablen. Die Hinzunahme von quadratischen Mischtermen zeigte beispielsweise keinen zusätzlichen Gewinn an Vorhersagegüte. Die gezeigten Modellstudien entsprechen einem linearen Ansatz ($N_p = 1$) oder einem Ansatz fünfter Ordnung ($N_p = 5$). Mit ersterem kann die Wichtigkeit der unabhängigen Variablen adäquat quantifiziert werden. Mit letzterem zeigte sich in vielen Tests eine (leicht) höhere Vorhersagegüte als mit dem linearen Ansatz.

Detaillierte Analyse der verschiedenen Modellstudien

Modellstudien basierend auf Umgebungsvariablen

Der Anstieg der Schärfe der Vorhersagen durch das *Resampling* lässt sich gut mittels U2 erkennen (Abbildung 6.13). Während die mittlere Ensemblevorhersage ohne vorheriges *Resampling* etwa zwischen 12 und 24 min liegt (Abbildung D.7) und damit die analysierten Unterschiede in Kapitel 5.3.2 (Abbildung 5.20a) widerspiegelt, variiert sie mit *Resampling*



(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

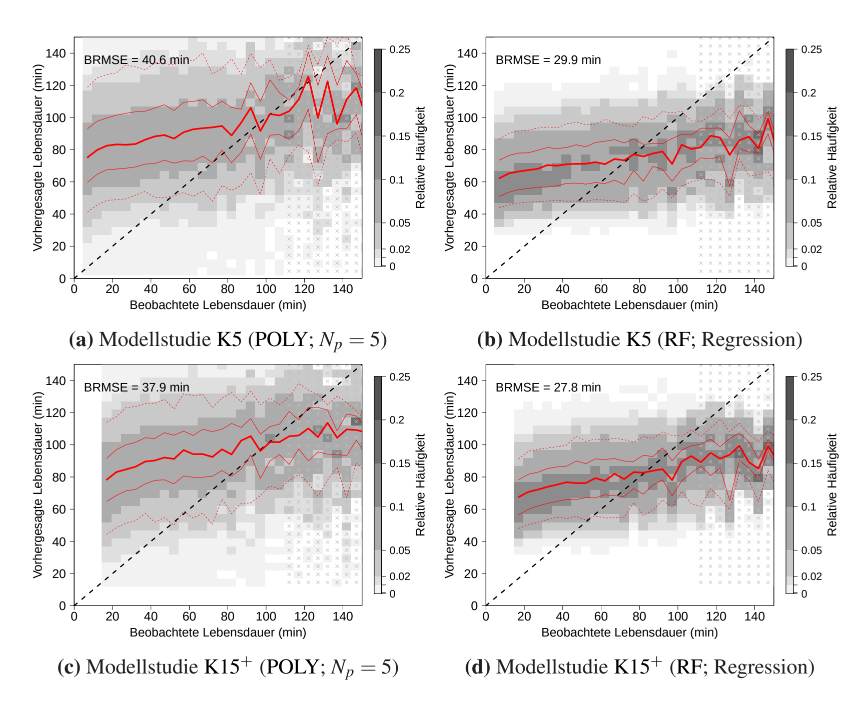
Abbildung 6.13: Mittlere Ensemblevorhersage für die Lebensdauer (min; Farbskala), aufgeteilt in verschiedene Gruppen der Prädiktoren DLS und LI_{100hPa}, für (a) den linearen Polynomansatz und (b) den *Random Forest* in U2 mit *Resampling*. Graue Kreuze geben Gruppen an, in denen 20 oder weniger Zellobjekte vorliegen.

zwischen etwa 40 und 120 (Polynomansatz) bzw. 100 min (*Random Forest*). Während die Vorhersageverfahren ohne *Resampling* demnach für Zellobjekte mit einer langen Lebensdauer eine viel zu kurze Lebensdauer prognostizieren, sagen sie mit *Resampling* für solche mit einer kurzen Lebensdauer eine viel zu lange Lebensdauer vorher (Abbildungen D.8a+b). Das Unterscheidungsvermögen der Vorhersagen ist auch mit *Resampling* weiterhin gering. Ein leichter Anstieg des Unterscheidungsvermögens zeigt sich bei der Hinzunahme weiterer Umgebungsvariablen in U6 und insbesondere in U15. Bei letzterer beträgt die Differenz des Medians der Ensemblevorhersagen für Zellobjekte mit einer Lebensdauer zwischen 7 und 107 min etwa 15 – 20 min (Abbildungen D.8c-f). Die Werte des *Balanced Root Mean Squared Errors* (*BRMSE* mit optimalem Wert 0; vgl. Kapitel 6.1.2) sinken jeweils leicht mit zunehmender Anzahl von Umgebungsvariablen. Sie sollten jedoch nicht zwischen den Vorhersageverfahren verglichen werden, da der *BRMSE* vom vorhergesagten Wertebereich beeinflusst ist, welcher wiederum sehr sensitiv und verfahrensabhängig auf

das *Resampling* reagiert. Je niedriger ϕ_{USP} ist, desto größer wird der Anteil von Zellobjekten mit eher längerer Lebensdauer, sodass der Wertebereich der Vorhersagen insgesamt hin zu höheren Werten verschoben wird. Der *Random Forest* erzielt generell deutlich niedrigere Werte für den *BRMSE* einzig durch den Umstand, dass seine Vorhersagen bei gleichem *Resampling* im Median etwa 10 min niedriger sind als die des Polynomansatzes.

Modellstudien basierend auf Zellattributen oder einer Kombination von Umgebungsvariablen und Zellattributen

Vorhersagen, die mit verschiedenen Kombinationen von Zellattributen zu Beginn der Zellentwicklung als Prädiktoren getroffen werden, zeigen ebenfalls ein maximales Unterscheidungsvermögen von etwa 15 – 20 min (Abbildung D.9). Noch etwa 5 min mehr lassen sich durch eine Kombination der 15 Umgebungsvariablen mit den entsprechenden Zellattributen erreichen (Abbildung 6.14). Der Polynomansatz mit nicht-linearen Termen bis zur fünften Ordnung ($N_p = 5$) schneidet dabei noch etwas besser ab als der *Random Forest*, welcher jedoch auch hier wiederum niedrigere (bessere) Werte für den *BRMSE* erzielt (s. o.). Die Vorhersagen zum Zeitpunkt 15 min nach der ersten Detektion der Zellobjekte durch KONRAD weisen niedrigere Werte für den *BRMSE* auf als die 5 min nach der ersten Detektion. Allerdings ist die Abnahme hauptsächlich auf das Wegfallen der Intervalle für eine beobachtete Lebensdauer von 7 bzw. 12 min zurückzuführen.
Verlässliche Vorhersagen können auch in diesen Modellstudien nur bedingt getroffen werden (Abbildung 6.15): Die Interquartilsbereiche der beobachteten Lebensdauer (Bereich zwischen den mitteldicken blauen Linien) nach unterschiedlichen Vorhersagen für die Lebensdauer überlappen sich stark. Der bedingte Median der beobachteten Lebensdauer (dicke blaue Linie) liegt meist in der Nähe der mittleren Lebensdauer der Zellobjekte des gesamten Datensatzes von rund 17 bzw. 31 min, wenn wie in K15$^+$ nur Objekte mit einer Lebensdauer von mehr als 15 min eingehen (niedrige Auflösung). Lediglich das Ensemble des *Random Forests* in K15$^+$ zeigt eine etwas bessere



(a) Modellstudie K5 (POLY; $N_p = 5$)

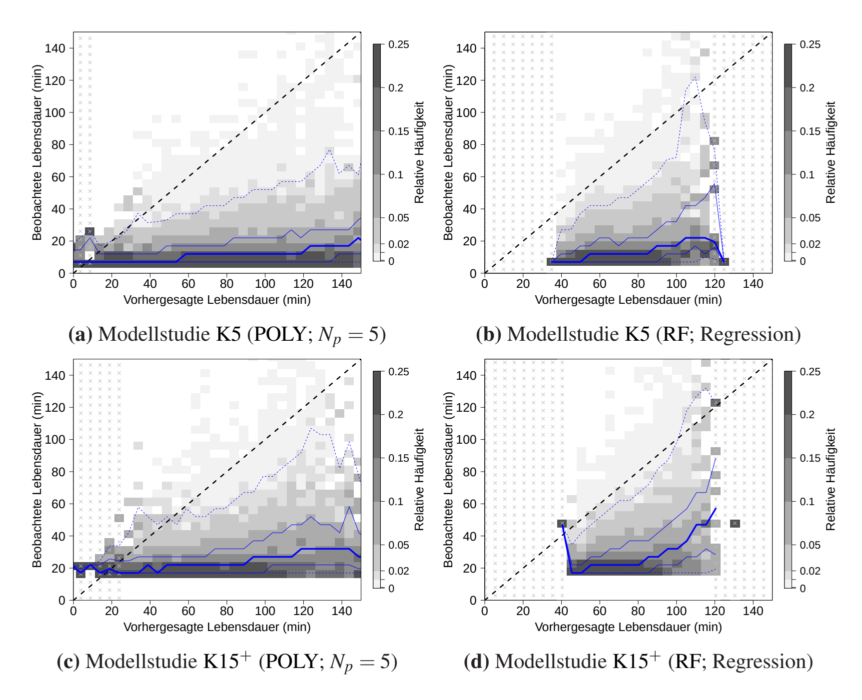
(b) Modellstudie K5 (RF; Regression)

(c) Modellstudie K15$^+$ (POLY; $N_p = 5$)

(d) Modellstudie K15$^+$ (RF; Regression)

Abbildung 6.14: Bedingte Histogramme und Quantil-Plots (*Likelihood-Base Rate Factorization*) basierend auf 51 Realisierungen (a,c) des Polynomansatzes mit $N_p = 5$ und (b,d) des *Random Forests* in (a,b) K5 bzw. (c,d) K15$^+$. Für jede beobachtete Lebensdauer ist die bedingte relative Häufigkeit der Lebensdauer-Vorhersagen in Graustufen dargestellt, d. h. die Häufigkeiten addieren sich in jeder Spalte zu 1 auf. Bereiche von beobachteten Werten für die Lebensdauer, die von weniger als 20 Zellobjekten vertreten werden, sind durch graue Kreuze markiert. Der Median ist als dicke rote Linie, die Werte für das 25. und 75. Perzentil sind als dünne rote Linien, und die für das 5. und 95. Perzentil als gestrichelte rote Linien eingetragen. Eine perfekte, deterministische Vorhersage würde der Diagonalen folgen. Links oben sind jeweils die Werte des *BRMSE* ergänzt.

Auflösung. Sagt das Ensemble im Mittel eine lange Lebensdauer von mehr als 100 min voraus, so ist folglich deutlich wahrscheinlicher eine mittellange oder lange Lebensdauer zu erwarten als nach der Vorhersage einer kurzen Lebensdauer. Generell ist jedoch in beiden Studien ein starkes *Overforecasting* zu erkennen, da der bedingte Median der Beobachtungen deutlich niedriger als die vorhergesagten Ensemblemittelwerte für die Lebensdauer ist.



(a) Modellstudie K5 (POLY; $N_p = 5$)

(b) Modellstudie K5 (RF; Regression)

(c) Modellstudie K15$^+$ (POLY; $N_p = 5$)

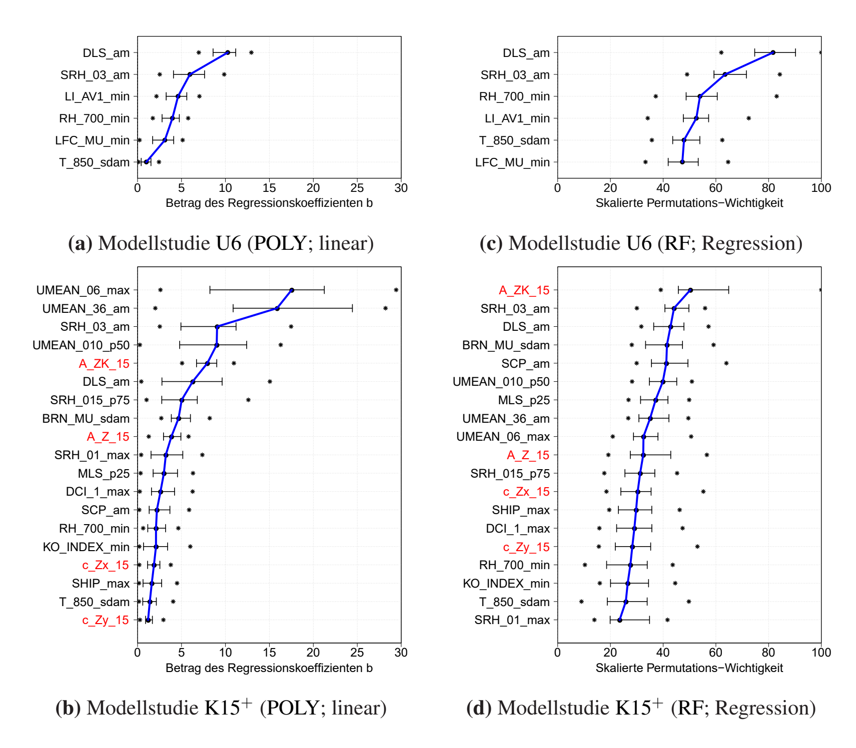
(d) Modellstudie K15$^+$ (RF; Regression)

Abbildung 6.15: Bedingte Histogramme und Quantil-Plots (*Calibration-Refinement Factorization*) basierend auf 51 Realisierungen (a,c) des Polynomansatzes mit $N_p = 5$ und (b,d) des *Random Forests* in (a,b) K5 bzw. (c,d) K15$^+$. Für jede vorhergesagte Lebensdauer (Ensemblemittel) ist die bedingte relative Häufigkeit der Lebensdauer-Beobachtungen in Graustufen dargestellt, d. h. die Häufigkeiten addieren sich in jeder Spalte zu 1 auf. Sonst analog zu Abbildung 6.14.

Zusammenfassende Analyse und Interpretation der Ergebnisse

Die kontinuierliche Vorhersage der Lebensdauer stellt sich mit den multivariaten Verfahren als schwierig heraus. Das maximale Unterscheidungsvermögen der Vorhersagen liegt bei etwa 25 min, d. h. die Vorhersagen der Verfahren liegen für Zellen mit einer kurzen Lebensdauer von nur etwa einer Viertelstunde im Median etwa 25 min niedriger als die Vorhersagen für Zellen, die anschließend eine lange Lebensdauer von mehr als anderthalb Stunden

erreichten. In den gezeigten Studien tritt zudem ein stark ausgeprägtes *Overforecasting* auf, verlässliche Vorhersagen lassen sich nur bedingt treffen. Das Ensemble des *Random Forests* in der kombinierten Studie K15^{+} zeigt dabei die beste Auflösung.
Das Modellensemble der jeweiligen Vorhersageverfahren kann auch als multivariates Analyse-Tool verstanden werden, welches die univariaten Analysen der Lebensdauer bezogen auf verschiedene Zellattribute in Kapitel 5.3.1 (Abbildung 5.5) und die uni- und bivariaten Analysen bezogen auf verschiedene Umgebungsvariablen in Kapitel 5.3 erweitert. Aus Abbildung 6.15 lassen sich Eintrittswahrscheinlichkeiten für die Lebensdauer in Abhängigkeit von der jeweiligen Vorhersage des Ensemblemittels ablesen. Eine prognostische Anwendung könnte so aussehen, dass bei gegebenen N_{ens} Vorhersagen eines Modellensembles entweder eine probabilistische Vorhersage der Lebensdauer direkt anhand der einzelnen Ensemblevorhersagen getroffen wird, oder der Ensemblemittelwert bestimmt wird und im Anschluss die Eintrittswahrscheinlichkeiten aus Abbildung 6.15 als probabilistische Vorhersage dienen.
Der Blick auf die Wichtigkeit der Prädiktoren zeigt, dass auch in den Regressionsverfahren dynamische Umgebungsvariablen mehr Einfluss haben als thermodynamische (Abbildung 6.16). In U6, in dem aus jedem Cluster aus Abbildung 5.14b eine Variable als Prädiktor dient, sind sowohl für den Polynomansatz als auch für den *Random Forest* die DLS und die $\mathrm{SRH}_{0-3\mathrm{km}}$ am bedeutsamsten. In der Kombination der 15 am besten unterscheidenden Umgebungsvariablen (Tabelle 5.1) mit den Zellattributen zum Zeitpunkt 5 bzw. 15 min nach der ersten Detektion ist neben einigen eng miteinander zusammenhängenden dynamischen Variablen besonders die Fläche des Zellkerns $A_{Z,K}$ wichtig, für den *Random Forest* belegt sie sogar den ersten Rang der skalierten Permutations-Wichtigkeit, wie schon bei den Modellstudien zur Klassifikation (vgl. Abbildung 6.11e). Die Verlagerung der Zellobjekte spielt hingegen auch hier keine große Rolle.



(a) Modellstudie U6 (POLY; linear)

(c) Modellstudie U6 (RF; Regression)

(b) Modellstudie K15$^+$ (POLY; linear)

(d) Modellstudie K15$^+$ (RF; Regression)

Abbildung 6.16: Wie Abbildung 6.11, nur für den linearen Polynomansatz und den *Random Forest* (Regression) in (a)+(c) U6, und (b)+(d) K15$^+$.

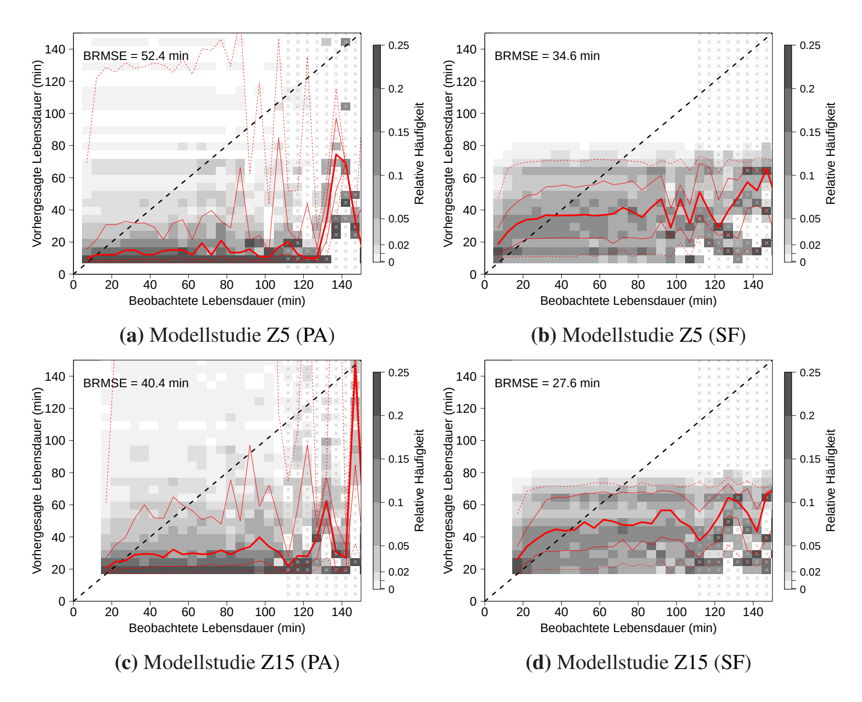
Einschub: Vorhersage mittels des Parabelansatzes und des Strömungsfeldansatzes

Vorhersage mittels des Parabelansatzes

Wie in den Kapiteln 5.1.2 und 5.3.1 erwähnt, kann der dort beschriebene Parabelansatz zur Vorhersage der zu erwartenden Lebensdauer sowie der maximalen Zellfläche angewendet werden. Ist zu einem bestimmten Zeitpunkt das Alter t (min) und die Fläche A_Z (km^2) eines Zellobjekts bekannt, kann die zu erwartende Lebensdauer deterministisch abgeschätzt und die Variabilität der Abschätzung über eine Fehlerrechnung quantifiziert werden (Anhang C).

Dabei bestätigt sich qualitativ die Folgerung aus Kapitel 5.1.2, dass sich alleine auf der Basis des vorgestellten Parabelansatzes eine (deterministische) Abschätzung der zu erwartenden Lebensdauer von konvektiven Zellen, insbesondere innerhalb der ersten halben Stunde, als eher schwierig erweist. Der Grund dafür ist, dass die Flächenentwicklungen der Zellobjekte mit einer unterschiedlich langen Lebensdauer sehr dicht beieinander liegen und kleine Abweichungen der Zellattribute einen großen Einfluss auf die verbleibende Lebensdauer haben (vgl. Abbildung 5.8a).
Wie sich dies quantitativ darstellt, wird mit der gleichen Vorgehensweise und demselben grundlegenden Algorithmus wie für die übrigen Vorhersageverfahren untersucht (vgl. Kapitel 6.1.1). Das Alter t und die korrespondierende Zellfläche der Zellobjekte $A_Z(t)$ dienen als Prädiktoren, welche keine Transformation erfahren. Jedes Ensemblemitglied bildet auf der Basis eines Trainingsdatensatzes – ohne Anwendung eines *Resamplings* – sein eigenes Parabelmodell, indem es die Koeffizienten μ_A und c_A bestimmt, wie in Kapitel 5.1.2 beschrieben. Anschließend findet Gleichung (C.1) Verwendung, um die Lebensdauer T_Z für jedes Zellobjekt im jeweiligen Testdatensatz abzuschätzen. Die Modelle berechnen nur Vorhersagen, wenn $A_Z(t) \leq 0{,}99\, A_{Z,krit}(t)$ ist.

Es zeigt sich, dass die Vorhersagen in den Modellstudien basierend auf dem Parabelansatz (PA) die Lebensdauer vieler Zellobjekte stark unterschätzen (Abbildungen 6.17a+c). Zeitpunkte für die Vorhersage sind $t = 7$ und 17 min, also 5 bzw. 15 min nach der ersten Detektion der Zellen durch KONRAD (Z5 bzw. Z15). Zellobjekte mit einer kürzeren Lebensdauer werden a priori aussortiert (vgl. Kapitel 6.1.1). Die Werte für das 75. Perzentil der prognostizierten Lebensdauer steigen und die des *BRMSE* sinken mit späteren Zeitpunkten für die Vorhersage zwar, sodass die Vorhersagen für $t = 17$ min beispielsweise nur einen Wert von $BRMSE = 40{,}4$ min aufweisen (Abbildung 6.17c). Dennoch ist das Unterscheidungsvermögen der Vorhersagen, dargestellt durch die Variation des Medians der Vorhersagen,



(a) Modellstudie Z5 (PA) **(b)** Modellstudie Z5 (SF)

(c) Modellstudie Z15 (PA) **(d)** Modellstudie Z15 (SF)

Abbildung 6.17: Wie Abbildung 6.14, nur basierend auf 51 Realisierungen (a,c) des Parabelmodells und (b,d) des Modells basierend auf dem Strömungsfeldansatz. Als Prädiktor wird ausschließlich die Zellfläche zum Zeitpunkt der (a,b) zweiten ($A_Z(t = 7\,\mathrm{min})$; Z5) bzw. (c,d) vierten ($A_Z(t = 17\,\mathrm{min})$; Z15) Detektion verwendet.

mit ca. $10 - 15\,\mathrm{min}$ recht gering. Damit liegt dieses Intervall in einem ähnlichen Bereich wie das Unterscheidungsvermögen des Polynomansatzes und des *Random Forests* in U2 (s. o.). Auch die Schärfe der Vorhersagen ist überschaubar.

Das niedrige Unterscheidungsvermögen wird durch die oben diskutierte Eigenschaft begünstigt, dass die Kurven der Parabeln, die zu unterschiedlichen Werten für die Lebensdauer gehören, gerade zu Beginn der Zellentwicklung bedingt durch die Konstruktion des Ansatzes in einem recht engen Intervall zwischen μ_A und $A_{Z,krit}(t)$ und darin sehr dicht beieinander

liegen (vgl. Abbildung 5.8a; Kapitel 5.1.2). Die real beobachteten Entwicklungen der Zellfläche sind jedoch deutlich weniger glatt als die idealisierten Parabeln, sodass es durch das Modell bei kleinen Unterschieden von A_Z zu großen Unterschieden in der Vorhersage kommen kann (Tabelle C.1). Verlässliche Vorhersagen mit einer adäquaten Auflösung können daher nicht erwartet werden (Abbildung D.10).

Vorhersage mittels des Strömungsfeldansatzes

Die Anwendung des Strömungsfeldansatzes zur quantitativen Untersuchung der Vorhersagegüte erfolgt mit der gleichen Vorgehensweise und demselben grundlegenden Algorithmus wie für die übrigen Vorhersageverfahren (vgl. Kapitel 6.1.1). Das Alter t und die korrespondierende Zellfläche $A_Z(t)$ der Zellobjekte dienen wie bei der Untersuchung des Parabelansatzes als Prädiktoren, welche keine Transformation erfahren. Jedes Ensemblemitglied bildet auf der Basis eines Trainingsdatensatzes – ohne Anwendung eines *Resamplings* – sein eigenes Strömungsfeld, wie in den Gleichungen (5.5) – (5.7) dargestellt ist (Kapitel 5.1.2). Anschließend findet Gleichung (5.8) Anwendung, um die Lebensdauer T_Z für jedes Zellobjekt im jeweiligen Testdatensatz über den Verlauf der entsprechenden Stromlinie abzuschätzen.

Es zeigt sich, dass die Vorhersagen in den Modellstudien basierend auf dem Strömungsfeldansatz (SF; Abbildungen 6.17b+d) die Lebensdauer der Zellobjekte besser abschätzen als diejenigen, die auf dem Parabelansatz basieren. Die Werte des *BRMSE* liegen um 10 – 15 min niedriger. Während die Modelle die Lebensdauer der Zellobjekte sogar leicht überschätzen, die weniger als etwa 40 min von KONRAD beobachtet wurden, prognostizieren sie Objekten mit einer längeren Lebensdauer meist eine zu kurze Lebensdauer. Das Unterscheidungsvermögen der Vorhersagen ist mit ca. 20 – 25 min höher als die der Parabelmodelle und liegt im Bereich des Polynomansatzes und des *Random Forests* in K15$^+$. Gerade Vorhersagen für Zellobjekte mit einer langen Lebensdauer von mehr als 40 min können

jedoch kaum unterschieden werden. Ähnlich wie für die Vorhersagen des *Random Forest*-Ensembles lässt sich dem Strömungsfeldansatz dennoch zumindest eine gewisse Auflösung attestieren (Abbildung D.10). Die längste vorhergesagte Lebensdauer liegt jedoch gerade einmal bei etwa $T_Z = 80\,\text{min}$. Dies lässt sich auf die geringere Anzahl von Zellobjekten in den jeweiligen Trainingsdatensätzen $f_{Tr}N$ der Ensemblemitglieder im Vergleich zur Analyse des gesamten Datensatzes zurückführen. Die geringere Anzahl führt dazu, dass durchgehende Stromlinien für einen weniger großen Bereich im Zellalter-Zellfläche-Raum $\mathscr{X}$ als für den gesamten Datensatz existieren, in dem durchgehende Stromlinien bis etwa $T_Z = 130\,\text{min}$ berechnet werden können (vgl. Abbildung 5.9).

Ausblick: Berücksichtigung von Umgebungsvariablen im Parabelansatz und dem Strömungsfeldansatz

Eine Verfeinerung des Parabelansatzes durch die Berücksichtigung einer Umgebungsvariablen wie z. B. dem LI könnte gewinnbringend sein (Kapitel 5.3.1 und Anhang C). Eine Erweiterung des Parabelmodells auf mehr als eine Umgebungsvariable in der vorgestellten Art und Weise stellt keine reale Option dar, da der Datensatz mit 38 553 Zellobjekten – davon aber nur 1 096 mit einer Lebensdauer von mehr als 60 min – dafür nicht ausreichend ist. Quantitative Modellstudien der Variante mit einer Umgebungsvariablen wurden bislang nicht durchgeführt. Der vorgestellte Ansatz zur Integration einer Umgebungsvariablen ist technisch sehr simpel in der Handhabung und der DWD könnte diesen mit wenig Aufwand in bereits bestehende Verfahren einarbeiten. Einzig die Bestimmung der Werte der Umgebungsvariablen erfordert eine Verbindung des *Nowcasting*-Verfahrens zu Feldern, welche die NWV liefert. Die Modellstudie SF wurde bereits unter Berücksichtigung des LI untersucht (nicht gezeigt). Eine solche Untersuchung kann beispielsweise auf zwei separaten Strömungsfeldern für $\text{LI}_{100\text{hPa}} < -1\,\text{K}$ und $\text{LI}_{100\text{hPa}} \geq -1\,\text{K}$ basieren, welche je nach beobachtetem Wert des $\text{LI}_{100\text{hPa}}$ eines Zellobjekts im

Testdatensatz zur Vorhersage der Lebensdauer Verwendung finden. Für Zellobjekte mit einer kurzen Lebensdauer von weniger als etwa 40 min führt dies zu einer leichten Verbesserung der Vorhersage, während weniger häufig hohe Werte für die Lebensdauer vorhergesagt werden. Der *BRMSE* liegt in diesem Fall sogar höher als in der Untersuchung mit nur einem Strömungsfeld pro Ensemblemitglied unabhängig vom $\mathrm{LI}_{100\mathrm{hPa}}$. Dies lässt sich damit erklären, dass durch die Aufteilung in zwei Strömungsfelder noch weniger Zellobjekte zur Erstellung des Strömungsfelds $\mathbf{v}_Z$ beitragen können. Somit liegen Tendenzen für höhere Zellalter nur sporadisch vor, sodass selten mehr als zehn Objekte zu einem Teilgebiet des Zellalter-Zellfläche-Raums $\mathscr{X}$ beitragen. Die Stromlinien reichen folglich erst gar nicht zu hohen Werten für die Lebensdauer. Beispielsweise reichen die Vorhersagen im Fall $\mathrm{LI}_{100\mathrm{hPa}} < -1\,\mathrm{K}$ nur bis etwa $T_Z = 70\,\mathrm{min}$, sodass hohe Werte für die Lebensdauer nicht prognostiziert werden. Ein solcher Ansatz lässt sich demnach nur mit einer größeren Stichprobe adäquat testen und anwenden, insbesondere mit einer deutlich höheren Anzahl von Zellobjekten mit einer langen Lebensdauer.

6.4 Modellstudien zur Vorhersage der maximalen Zellfläche

6.4.1 Evaluation von Klassifikationsverfahren zur Vorhersage der maximalen Zellfläche

Analog zu den Modellstudien zur Untersuchung der Ensemblevorhersagen für den binären Prädiktanden kurze/lange Lebensdauer von Zellobjekten (Kapitel 6.3.1) folgt eine Auswertung ähnlicher Studien zur Vorhersage der zu erwartenden maximalen Zellfläche $A_{Z,max}$ (Tabelle 6.3). Nach der Diskussion der Vorhersagen für die Lebensdauer in den vorigen Abschnitten beschränken sich die folgenden Untersuchungen auf eine Auswahl von fünf Studien, anhand derer die wesentlichen Aspekte der Evaluation präsentiert werden.

Der Klassentrennwert, der zwischen räumlich begrenzten (kleinen) und ausgedehnten (großen) Zellobjekten unterscheidet, ist $\chi = 60\,\mathrm{km}^2$ mit einer halben Breite des Übergangsbereichs von $\chi' = 10\,\mathrm{km}^2$, sodass Zellobjekte mit einer maximalen Zellfläche zwischen $\chi - \chi' = 50\,\mathrm{km}^2$ und $\chi + \chi' = 70\,\mathrm{km}^2$ nicht in die Berechnung der Gütemaße eingehen. Durch diese Wahl von χ beträgt das Klassenverhältnis $\rho_K \approx 2{,}7\,\%$ – ähnlich wie in den Untersuchungen zur Lebensdauer. Auch die Anzahl von evaluierten kleinen und großen Zellobjekten ist durch die Wahl von χ' ähnlich wie in den genannten Untersuchungen. Erfolgt die Anwendung eines *Resamplings*, so geschieht dies für die folgenden Studien in Bezug auf die Verteilung der maximalen Zellfläche (vgl. Kapitel 3.5.2 und 6.1.1).

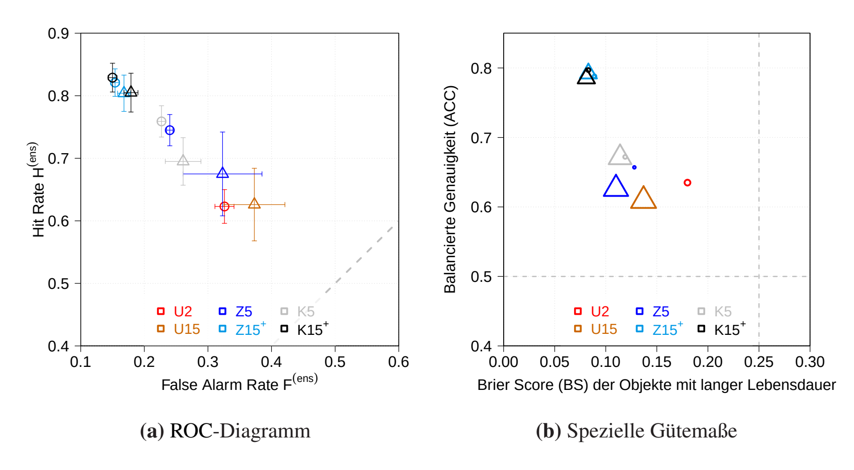
Detaillierte Analyse der verschiedenen Modellstudien

Die Vorhersagen der maximalen Zellfläche, die nur mit Umgebungsvariablen als Prädiktoren getroffen werden, erreichen leicht bessere Werte für verschiedene Gütemaße als analoge Vorhersagen der Lebensdauer (Abbildung 6.18; vgl. Abbildung 6.9). In den Untersuchungen zum Unterscheidungsvermögen der Umgebungsvariablen hinsichtlich der Lebensdauer und der maximalen Zellfläche (Kapitel 5.3.1) erzielten die Umgebungsvariablen bei der Lebensdauer bessere Gütemaße als bei der maximalen Zellfläche, was zum Teil mit der Wahl der dortigen Klassentrennwerte zusammenhängt. Darüber hinaus können nicht-lineare und schwer nachvollziehbare Ursachen die Vorhersagen beeinflussen, sodass die Gütemaße der Vorhersagen für die maximale Zellfläche hier leicht bessere Werte erzielen. In U2 mit der logistischen Regression erreichen die DLS und der $\mathrm{LI}_{100\mathrm{hPa}}$ einander ähnliche Effekt-Koeffizienten, und in U15 mit dem *Random Forest* sind dynamische und thermodynamische Variablen etwa gleich wichtig (nicht gezeigt). Vorhersagen der maximalen Zellfläche, die nur mit Zellattributen zu Beginn der Zellentwicklung als Prädiktoren getroffen werden, erzielen deutlich bessere Werte für verschiedene

Tabelle 6.3: Übersicht über die verschiedenen Kombinationen von Prädiktoren sowie die verwendeten Entscheidungstrennwerte μ für unterschiedliche Modellstudien zur Vorhersage der maximalen Zellfläche. Oben: Modellstudien, die entweder nur auf Umgebungsvariablen oder nur auf Zellattributen basieren; unten: Modellstudien, die auf einer Kombination von Umgebungsvariablen und Zellattributen basieren. Alle Werte der Umgebungsvariablen entsprechen denjenigen zum Zeitpunkt der ersten Detektion der Zellen durch KONRAD.

Modellstudie →	U2	U15	Z5	Z15$^+$
Prädiktoren	DLS, $\mathrm{LI}_{100\mathrm{hPa}}$	15 beste Variablen aus Tabelle 5.2	A_Z, $A_{Z,K}$ ($t = 7$ min)	A_Z, $A_{Z,K}$, $c_{Z,x}$, $c_{Z,y}$ ($t = 17$ min)
μ_{LOGR}	0,029	—	0,029	0,075
μ_{RF}	—	0,450	0,150	0,050
Resampling für RF	—	✓	✓	×

Modellstudie →	K5	K15$^+$	K15$^+_{var}$
Prädiktoren (LOGR)	wie U2 + Z5	wie U2 + Z15$^+$	wie K15$^+$
Prädiktoren (RF)	wie U15 + Z5	wie U15 + Z15$^+$	wie K15$^+$
μ_{LOGR}	0,029	0,075	[0,01 ; 0,21]
μ_{RF}	0,420	0,070	[0,01 ; 0,21]
Resampling für RF	✓	×	×



(a) ROC-Diagramm **(b)** Spezielle Gütemaße

Abbildung 6.18: Analog zu Abbildung 6.9, nur mit der maximalen Zellfläche als Prädiktand.

Gütemaße als analoge Vorhersagen der Lebensdauer. Insbesondere ist bemerkenswert, dass sich die maximale Zellfläche zum Zeitpunkt 15 min nach der ersten Detektion durch KONRAD allein auf der Basis der Zellattribute viel besser abschätzen lässt als die Lebensdauer durch eine Kombination von Zellattributen und Umgebungsvariablen. Auch in Z5 erreichen die Gütemaße zur Vorhersage der Zellfläche höhere Werte als für die Lebensdauer. Hier profitieren die Vorhersagen zudem von der Information über die vorherrschenden Umgebungsbedingungen (K5), während sie schon zum Zeitpunkt 15 min nach der ersten Detektion kaum mehr einen zusätzlichen Gewinn bringen. Zu Beginn der Zellentwicklung sind daher die Umgebungsvariablen zur Abschätzung der maximalen Zellfläche nützlich, während im weiteren Verlauf fast ausschließlich die Zellhistorie relevant ist.

Die ROC-Kurven für Z5 und Z15$^+$ zeigen, dass beide Vorhersageverfahren gut zwischen kleinen Zellobjekten mit einer maximalen Zellfläche von weniger als 50 km^2 und großen Zellobjekten mit einer maximalen Zellfläche von mehr als 70 km^2 differenzieren können (Abbildung 6.19). Dabei erzielt die logistische Regression leicht höhere Werte für die *AUC*. Die Zuordnung

der großen Objekte ist zu mehr als 94 % korrekt, wenn man eine indifferente Trefferrate von nur 50 % für die kleinen Zellobjekte in Kauf nimmt (analog zu Fragestellung (A) aus Kapitel 6.2.2). Die ROC-Kurven verlaufen im Bereich niedriger F-Werte sehr steil, sodass für die logistische Regression (den *Random Forest*) bei $F \approx 0{,}07$ etwas mehr (weniger) als drei Viertel aller großen Zellobjekte korrekt abgeschätzt werden (analog zu Fragestellung (C)). Für große Trennwerte μ sind damit trotz des Ungleichgewichts des Datensatzes bezüglich der maximalen Zellfläche sogar Werte für das Fehlalarmverhältnis FAR von deutlich weniger als 0,5 bei gleichzeitigem Erreichen von $H > 0{,}5$ möglich. Durch die Erweiterung des Wertebereichs der Entscheidungstrennwerte im Vergleich zur Modellstudie K15^{+} ergibt sich für jeweils ein exemplarisches Modell beispielsweise:

$$\begin{aligned}
\mu_{RF} = 0{,}210 \quad &\Longrightarrow \quad FAR = 0{,}365\,; \quad H = 0{,}641 \\
\mu_{RF} = 0{,}282 \quad &\Longrightarrow \quad FAR = 0{,}153\,; \quad H = 0{,}595 \\
\mu_{LR} = 0{,}330 \quad &\Longrightarrow \quad FAR = 0{,}128\,; \quad H = 0{,}607\,. \qquad (6.8)
\end{aligned}$$

Dies ist bei den Vorhersagen der Lebensdauer nicht realisierbar.

Zusammenfassende Analyse und Interpretation der Ergebnisse

Die Abschätzung einer binären Klasse für die maximale Zellfläche zur Unterscheidung von räumlich wenig und weit ausgedehnten Zellobjekten erreicht mit beiden Klassifikationsverfahren teilweise wesentlich bessere Gütemaße als für die Lebensdauer. Im Vergleich zur Klassifikation der Lebensdauer ist die Abschätzung der maximalen Zellfläche deutlich von der Zellfläche A_Z zu Beginn der Zellentwicklung bestimmt, d. h. die relative Wichtigkeit von Umgebungsvariablen als Prädiktoren in Kombination mit der Information über die Zellfläche in einem Ensemble der logistischen Regression oder des *Random Forests* ist geringer (Abbildung 6.20). Dies deckt sich mit den Analysen zum Unterscheidungsvermögen der Variablen hinsichtlich der maximalen

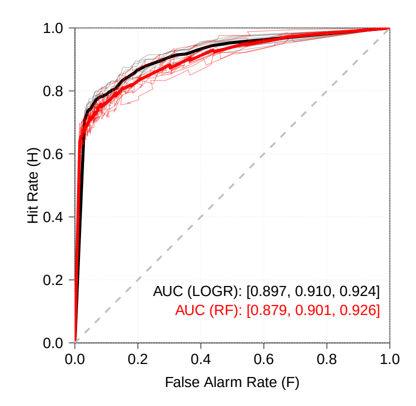
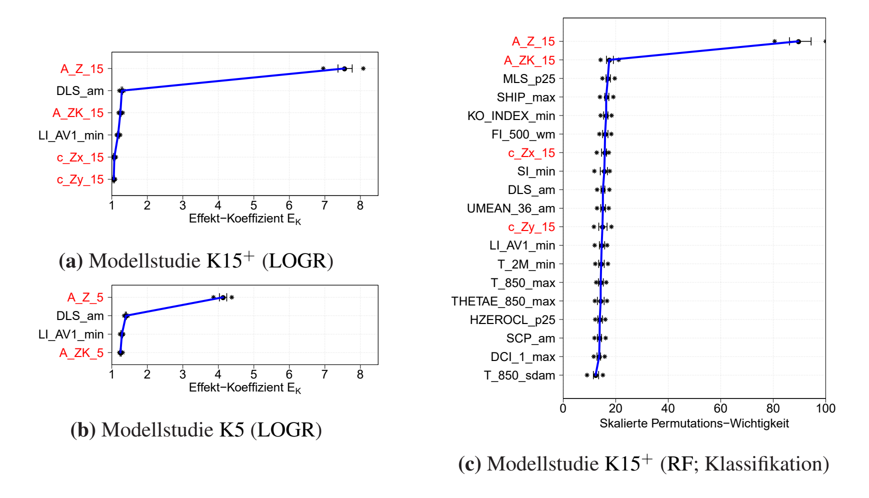


Abbildung 6.19: Wie Abbildung 6.12, nur mit der maximalen Zellfläche als Prädiktand.



(a) Modellstudie K15$^+$ (LOGR)

(b) Modellstudie K5 (LOGR)

(c) Modellstudie K15$^+$ (RF; Klassifikation)

Abbildung 6.20: Wie Abbildung 6.11, nur mit der maximalen Zellfläche als Prädiktand in (a)+(c) K15$^+$ und (b) K5 (LOGR).

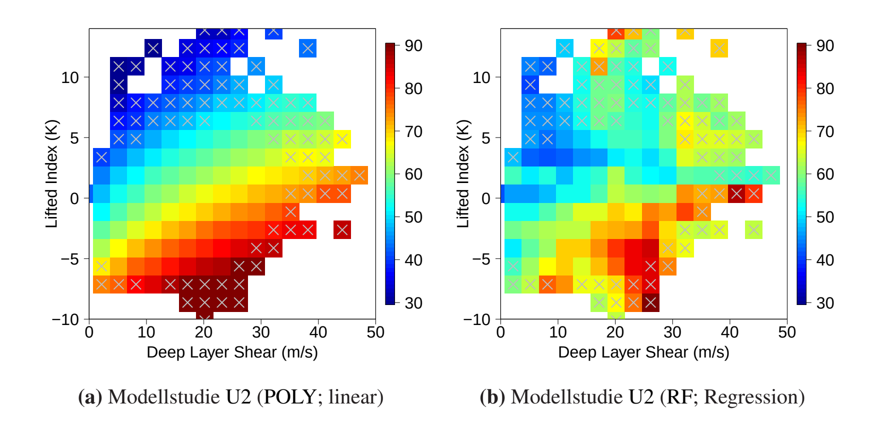
Zellfläche (vgl. Kapitel 5.3.1). Des Weiteren bedeutet dieses Ergebnis, dass insbesondere die Entwicklung der Fläche einer konvektiven Zelle zu Beginn ihres Lebenszyklus bereits auf ihre maximale Zellfläche hindeutet, die sie im

weiteren Verlauf erreichen wird (vgl. Kapitel 5.1.2). Dennoch verbessert die Information über die Umgebungsvariablen die Prognosen in der ersten Viertelstunde der Zellentwicklung merklich, insbesondere für ein Ensemble des *Random Forests*, während im weiteren Verlauf fast ausschließlich die Zellhistorie relevant ist.

6.4.2 Evaluation von Regressionsverfahren zur Vorhersage der maximalen Zellfläche

Detaillierte Analyse der verschiedenen Modellstudien

Analog zu Kapitel 6.3.2 werden nachfolgend dieselben fünf Modellstudien wie für die Klassifikationsverfahren mit zwei Klassen der maximalen Zellfläche $A_{Z,max}$ als Prädiktand vorgestellt (vgl. Tabelle 6.3). Auch hier lässt sich der Anstieg der Schärfe der Vorhersagen durch das *Resampling* gut mittels U2 erkennen (Abbildung 6.21). Während die mittlere Ensemblevorhersage ohne vorheriges *Resampling* etwa zwischen 23 und 35 km^2 liegt (Abbildung D.7) und damit die analysierten Unterschiede in Kapitel 5.3.2 (Abbildung 5.22a) widerspiegelt, variiert sie mit *Resampling* zwischen etwa 30 und 90 (Polynomansatz) bzw. 80 km^2 (*Random Forest*). Ohne *Resampling* prognostizieren die Modelle Objekten mit einer großen Zellfläche meist eine zu kleine Fläche, mit *Resampling* sagen sie Objekten mit einer kleinen Zellfläche eine zu große Fläche vorher (Abbildungen D.12a+b). Wie bei der Vorhersage der Lebensdauer ist das Unterscheidungsvermögen der Vorhersagen auch mit *Resampling* weiterhin gering. Die Hinzunahme weiterer Umgebungsbedingungen als Prädiktoren führt nur zu einem sehr geringen Anstieg des Unterscheidungsvermögens, wie sich in U6 und U15 zeigt (Abbildungen D.12c-f). Die Werte des *BRMSE* sinken jeweils kaum mit zunehmender Anzahl von Umgebungsvariablen.



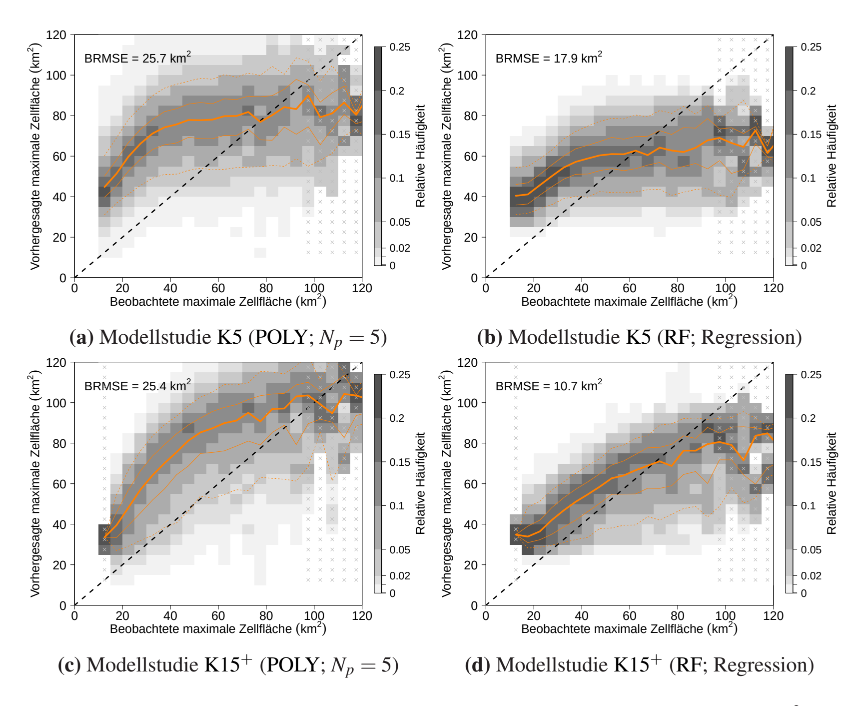
(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

Abbildung 6.21: Wie Abbildung 6.13, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2; Farbskala) als Prädiktand.

Vorhersagen, die mit verschiedenen Kombinationen von Zellattributen zu Beginn des Lebenszyklus als Prädiktoren getroffen werden, zeigen ein deutlich besseres Unterscheidungsvermögen (Abbildung D.13). Die Modelle prognostizieren Zellobjekten mit einer kleinen Zellfläche auch hier zu große Zellflächen. Diese fallen jedoch insbesondere für die häufig auftretenden, sehr kleinen Zellobjekte deutlich kleiner aus als in den Studien mit den Umgebungsvariablen, insbesondere in den Untersuchungen mit dem *Random Forest*. Die Vorhersagen in Z15$^+$ sind für beide Vorhersageverfahren nochmals besser als in Z5, d. h. 15 min nach der ersten Detektion einer Zelle durch KONRAD unterscheiden sich die Entwicklungen von klein bleibenden und groß anwachsenden Zellobjekten deutlich.

Das beste Unterscheidungsvermögen lässt sich auch hier durch eine Kombination der 15 Umgebungsvariablen mit den entsprechenden Zellattributen erreichen, wenn auch die Umgebungsvariablen nur einen geringen



(a) Modellstudie K5 (POLY; $N_p = 5$)

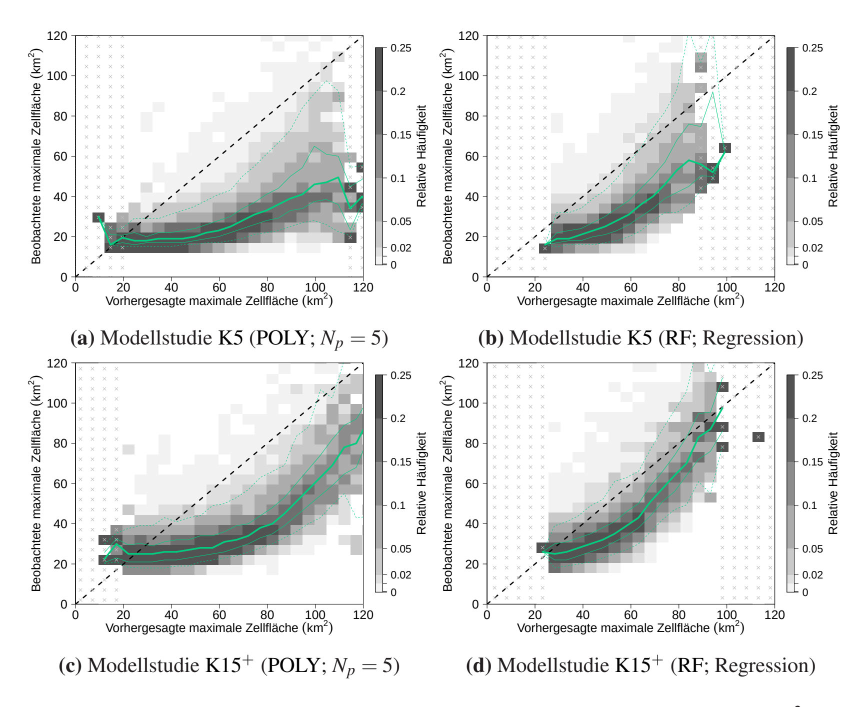
(b) Modellstudie K5 (RF; Regression)

(c) Modellstudie K15$^+$ (POLY; $N_p = 5$)

(d) Modellstudie K15$^+$ (RF; Regression)

Abbildung 6.22: Wie Abbildung 6.14, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.

Einfluss haben, wie die ähnlichen Werte des *BRMSE* andeuten (Abbildung 6.22; s. u.). Der Polynomansatz mit nicht-linearen Termen bis zur fünften Ordnung ($N_p = 5$) weist dabei eine höhere Schärfe als der *Random Forest* auf, welcher dafür auch hier wiederum niedrigere (bessere) Werte für den *BRMSE* erzielt. Auch ein linearer Polynomansatz mit $N_p = 1$ führt bereits zu sehr ähnlichen Vorhersagen (nicht gezeigt). Die Vorhersagen zum Zeitpunkt 15 min nach der ersten Detektion weisen niedrigere Werte für den *BRMSE* auf als die 5 min nach der ersten Detektion. Die Abnahme ist hier neben dem Wegfallen der Intervalle für eine beobachtete Lebensdauer von 7 bzw. 12 min auf eine deutliche Reduzierung der Vorhersagen von großen



(a) Modellstudie K5 (POLY; $N_p = 5$)

(b) Modellstudie K5 (RF; Regression)

(c) Modellstudie K15$^+$ (POLY; $N_p = 5$)

(d) Modellstudie K15$^+$ (RF; Regression)

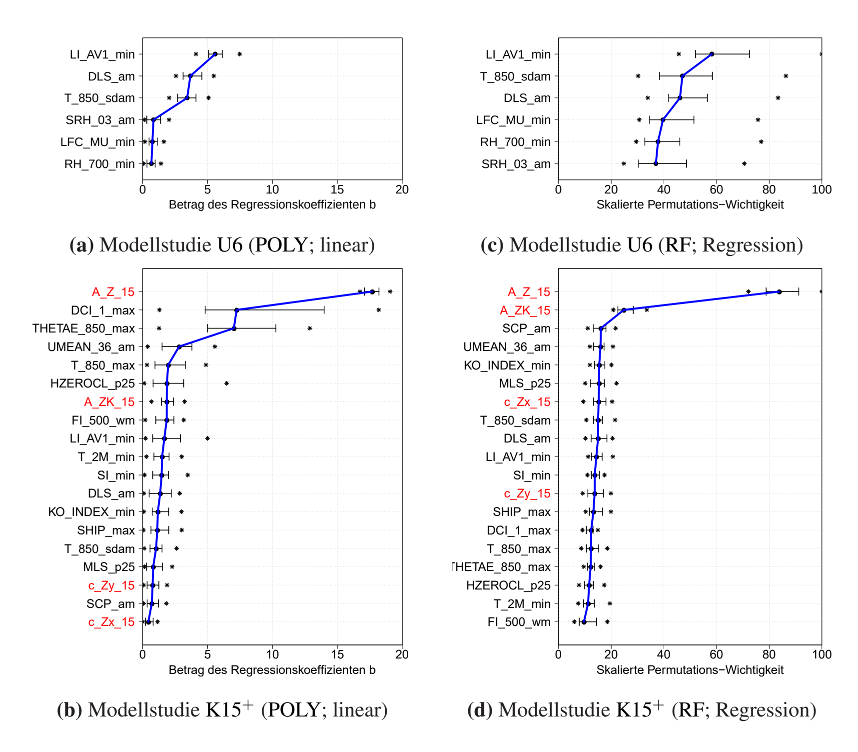
Abbildung 6.23: Wie Abbildung 6.15, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.

Zellflächen zurückzuführen. Verlässliche Vorhersagen für die maximale Zellfläche können im Gegensatz zu äquivalenten Vorhersagen der Lebensdauer unter Berücksichtigung des *Overforecastings* gut getroffen werden (Abbildung 6.23; vgl. Abbildung 6.15). Die gute Auflösung der Vorhersagemodelle ist an dem sichtbaren Anstieg der Beobachtungen entsprechend zu den jeweiligen Vorhersagen zu erkennen. Die Interquartilsbereiche und sogar die Bereiche zwischen dem 5. und dem 95. Perzentil der beobachteten maximalen Zellfläche (Bereiche zwischen den mitteldicken bzw. dünnen grünen Linien) nach unterschiedlichen Vorhersagen für die maximale Zellfläche überlappen sich in K15$^+$ deutlich weniger als für die Vorhersage der Lebensdauer. Der bedingte Median der beobachteten maximalen Zellfläche (dicke grüne Linie)

weicht teilweise deutlich vom Mittelwert der Zellfläche des gesamten Datensatzes von 28 bzw. 36 km^2 ab, wenn wie in $K15^+$ nur Objekte mit einer Lebensdauer von mehr als 15 min eingehen. Die Auflösung für die Vorhersagen der maximalen Zellfläche in der Modellstudie K5 ist sogar besser als die für die Vorhersagen der Lebensdauer in $K15^+$ (vgl. Abbildung 6.15). Die maximale Zellfläche kann demnach 5 min nach der ersten Detektion der Zelle bereits verlässlicher abgeschätzt werden als die Lebensdauer 15 min nach der ersten Detektion. Insbesondere mit dem Ensemble des *Random Forests* in $K15^+$ besteht eine hohe Verlässlichkeit der Vorhersagen sowohl für K5 als auch $K15^+$, d. h. die Eintrittswahrscheinlichkeit von Zellobjekten mit einer großen Zellfläche ist bei Vorhersagen größerer Zellflächen sehr hoch und bei solchen einer kleineren Zellfläche sehr gering.

Zusammenfassende Analyse und Interpretation der Ergebnisse

Die Schlussfolgerungen der Modellstudien mit den Klassifikationsverfahren lassen sich auf die Regressionsverfahren übertragen (vgl. Kapitel 6.4.1). Nach der ersten Detektion einer Zelle durch KONRAD unterscheiden sich die Entwicklungen von klein bleibenden und groß anwachsenden Zellobjekten deutlich, sodass verlässliche Vorhersagen getroffen werden können. Die maximale Zellfläche kann somit bereits 5 min nach der ersten Detektion der Zelle besser abgeschätzt werden als die Lebensdauer 10 min später. Der Blick auf die Wichtigkeit der Prädiktoren zeigt, dass auch in den Regressionsverfahren zur Vorhersage der maximalen Zellfläche thermodynamische und dynamische Umgebungsvariablen einen ähnlichen, wenn auch geringen Einfluss haben (Abbildung 6.24). In U6, in dem aus jedem Cluster aus Abbildung 5.14b eine Variable als Prädiktor dient, ist sowohl für den Polynomansatz als auch für den *Random Forest* der LI_{100hPa} am bedeutsamsten. In der Kombination der 15 am besten unterscheidenden Umgebungsvariablen aus Tabelle 5.2 mit den Zellattributen zum Zeitpunkt 5 bzw. 15 min nach der ersten Detektion nimmt die jeweilige Zellfläche A_Z in beiden Verfahren eine überragende



(a) Modellstudie U6 (POLY; linear)

(c) Modellstudie U6 (RF; Regression)

(b) Modellstudie K15$^+$ (POLY; linear)

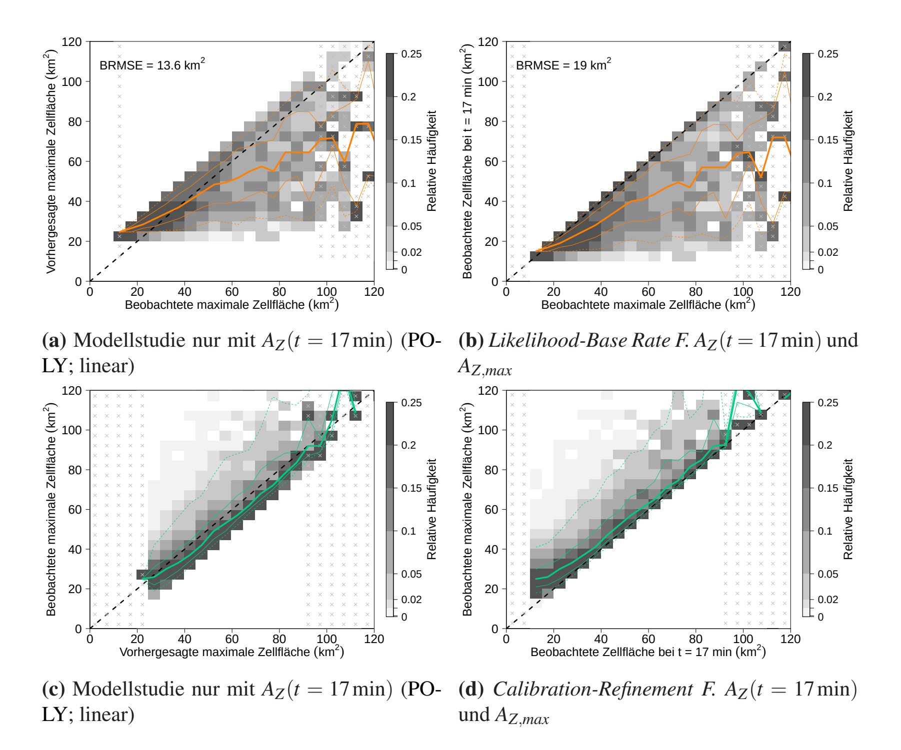
(d) Modellstudie K15$^+$ (RF; Regression)

Abbildung 6.24: Wie Abbildung 6.16, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.

Rolle ein – wie nach den obigen Untersuchungen und den Ergebnissen aus Kapitel 6.4.1 erwartet. Interessanterweise sind für den Polynomansatz auch zwei thermodynamische Variablen ($\text{DCI}_{100\text{hPa}}$ und $\theta_{\text{ps},850\text{hPa}}$) von Bedeutung. Die quantitative Vorhersage der maximalen Zellfläche mittels des Parabelansatzes und dem Strömungsfeldansatz aus Kapitel 5.1.2, in der gleichen Weise durchgeführt wie zur Untersuchung der Lebensdauer in Kapitel 6.3.2, liefert weniger gute Ergebnisse (Abbildungen D.15 und D.16). Da die Gründe hierfür dieselben wie für die Vorhersage der Lebensdauer sind, werden diese Studien hier nicht weiter diskutiert.

Einordnung der Vorhersagen aus den Modellstudien

Unter einer geringen Einbuße von Vorhersageschärfe können unter Verwendung eines schwächeren *Resamplings* mit $\phi_{USP} = 0{,}85$ in K15^{+} die Werte des *BRMSE* verringert werden (Abbildung D.14). Damit bleibt der Trainingsdatensatz in einem gewissen Maße unbalanciert. Erfolgt gar kein *Resampling*, so reduziert sich das *Overforecasting* der kleineren Zellobjekte, dafür stellt sich ein leichtes *Underforecasting* der größeren Zellobjekte ein (Abbildung D.17). Hier sagt das Ensemble des Polynomansatzes jedoch so gut wie keine Zellflächen von mehr als 65 km^2 mehr vorher. Das Ensemble des *Random Forests* weist zwar leicht höhere Werte des *BRMSE* auf als im Fall mit *Resampling*, die Verlässlichkeit der Vorhersagen ist ohne *Resampling* jedoch am höchsten (Abbildung D.17d).
Abschließend stellt sich die Frage, wie diese Vorhersagen im Vergleich zu einer vereinfachten Vorhersage einzuordnen sind. Die Korrelation der Zellfläche zu einem bestimmten Zeitpunkt zu Beginn des Lebenszyklus (z. B. in der ersten halben Stunde) und der maximalen Zellfläche liegt bei $r_P \in [0{,}68\,;0{,}74]$. Darüber hinaus erzielt bereits die Vorhersage, dass die maximale Zellfläche der Fläche zum Zeitpunkt der vierten Detektion entspricht, gute Ergebnisse (Abbildungen 6.25b+d). Viele Zellobjekte wachsen anschließend nur noch geringfügig (vgl. Kapitel 5.1.2). Es kommt dadurch insgesamt zu einem leichten *Underforecasting*. Ein linearer Polynomansatz ohne *Resampling*, Prädiktortransformation und Dämpfung (Regularisierungsparameter $\alpha = 0$; einfache lineare Regression, vgl. Kapitel 3.3.3) mit nur der Zellfläche A_Z zu einem bestimmten Zeitpunkt als Prädiktor führt somit ebenfalls bereits zu einer sehr guten Abschätzung (Abbildungen 6.25a+c), die in etwa der des *Random Forests* ohne *Resampling* entspricht. Die Hinzunahme weiterer Prädiktoren zu diesem einfachen Ansatz bringt daher lediglich einen sehr geringen Gewinn.



(a) Modellstudie nur mit $A_Z(t = 17\,\text{min})$ (POLY; linear)

(b) *Likelihood-Base Rate F.* $A_Z(t = 17\,\text{min})$ und $A_{Z,max}$

(c) Modellstudie nur mit $A_Z(t = 17\,\text{min})$ (POLY; linear)

(d) *Calibration-Refinement F.* $A_Z(t = 17\,\text{min})$ und $A_{Z,max}$

Abbildung 6.25: (a)+(c) Wie Abbildungen D.17a+c, nur mit einem linearen Ansatz, ohne Prädiktortransformation und ohne Dämpfung ($\alpha = 0$). (b)+(d) Einfache Gegenüberstellung der beobachteten Zellfläche zum Zeitpunkt 15 min nach der ersten Detektion und der maximalen Zellfläche, in (b) der *Likelihood-Base Rate Factorization* und (d) der *Calibration-Refinement Factorization*.

7 Zusammenfassung, Diskussion und Ausblick

Die vorliegende Arbeit stellte eine umfangreiche Analyse der Lebenszyklen konvektiver Zellen in Deutschland und ihren Zusammenhang zu den vorherrschenden Umgebungsbedingungen in der Atmosphäre vor. Basierend auf einem kombinierten, einzigartigen Datensatz aus Zellattributen und einer Vielzahl von konvektionsrelevanten meteorologischen Variablen wurden darüber hinaus verschiedene statistische Vorhersagemodelle zur Abschätzung der Lebensdauer und der Größe konvektiver Zellen entwickelt und mit dem Ziel untersucht, herauszufinden, welche von ihnen für eine Verbesserung von *Nowcasting*-Verfahren im automatischen Warnprozess des Deutschen Wetterdienstes (DWD) geeignet sind.

Auf der Basis von Radardaten konnte der operationelle Zellverfolgungsalgorithmus KONRAD des DWD Daten von konvektiven Zellen im Bundesgebiet und benachbarten Regionen generieren, welche die Grundlage für die Untersuchung von Lebenszyklen, also der zeitlichen Entwicklung verschiedener Zellattribute, bildeten. Diese Daten lagen für die Sommerhalbjahre 2011 – 2016 (April bis September) vor und wurden mit Hilfe einer meteorologisch fundierten Qualitätskontrolle und einer entsprechenden Filterung zu einem Datensatz von zusammenhängenden Lebenszyklen konvektiver Zellen (Zellobjekte) verarbeitet. Gleichzeitig standen zeitlich und räumlich hochaufgelöste Assimilationsanalysen des numerischen Wettervorhersagemodells COSMO-EU (DWD) zur Verfügung, welche die Berechnung vieler konvektionsrelevanter meteorologischer Variablen zur Charakterisierung

der atmosphärischen Umgebungsbedingungen ermöglichten. Hierzu wurden entsprechende Routinen des COSMO-Modells zur Ausgabe und Nachbereitung von Daten erweitert und die neuen Variablen mit dem Modell nachsimuliert. Die Implementierung der neu hinzugefügten Umgebungsvariablen wurde mit Hilfe von Radiosondendaten evaluiert.
Anschließend wurde der Datensatz der Lebenszyklen mit den Umgebungsvariablen kombiniert. Die Umsetzung erfolgte durch die Einführung eines adaptiven Umgebungsradius für die Zellobjekte, innerhalb dessen verschiedene statistische Eigenschaften der Umgebungsvariablen zum Zeitpunkt der Detektion der Zellen abgeschätzt wurden. Dadurch konnte ein einzigartiger, kombinierter objektbezogener Datensatz generiert werden. Die knapp 40 000 Lebenszyklen enthalten in fünfminütlicher Auflösung unter anderem Informationen über die zeitliche Entwicklung der Position, der Verlagerung und der Größe der Zellen sowie der Größe des Zellkerns, der den Bereich des intensivsten Niederschlags einer Zelle darstellt. Diese Informationen wurden hier direkt mit den Werten von mehr als 50 Umgebungsvariablen in Verbindung gebracht, wie beispielsweise mit Maßen für den Feuchtegehalt der Atmosphäre, die vertikale Windscherung und die thermische Stabilität.
Der kombinierte Datensatz stellte die Grundlagen für alle folgenden Analysen und Entwicklungen dar. Um die Charakteristika der Lebenszyklen besser zu verstehen und die Analysen tiefgreifend interpretieren zu können, wurden die Lebenszyklen zunächst unabhängig von den Umgebungsbedingungen untersucht. Die transiente und stochastische Natur konvektiver Zellen sowie die mathematisch-technischen Limitierungen des Zellverfolgungsalgorithmus KONRAD erforderten unter anderem das Aussortieren von einigen Zellobjekten, die auf der Basis der vorliegenden Daten keinen kompletten, repräsentativen Lebenszyklus darstellen konnten. Insbesondere Multizellen und Mesoskalige Konvektive Systeme werden durch KONRAD nicht als ein zusammenhängendes System erfasst. Die erstellten Lebenszyklen repräsentieren somit nur einen Teil des konvektiven Spektrums, nämlich isolierte Einzelzellen und Superzellen. Nichtsdestoweniger war

diese Auswahl im Hinblick auf die oben beschriebene Zielsetzung der vorliegenden Arbeit sehr gut geeignet, insbesondere, da für das *Nowcasting* eine Abschätzung der weiteren Entwicklung konvektiver Zellen unabhängig von ihrer Organisationsform von Interesse war.

Die Zellen zeigten eine typische, von den großräumigen Bedingungen bestimmte Verteilung ihrer Zugrichtung. Der Anteil der etwa von Südwest nach Nordost ziehenden Zellen war am größten. Bei der Untersuchung der KONRAD-Daten wurde auch deutlich, dass die meisten Zellen eine recht kurze Lebensdauer haben und dementsprechend eine kurze Zugbahn und kleine flächenhafte Ausdehnung aufweisen. MacKeen et al. (1999) illustrierten bereits, dass diese Tatsache eine Prognose der verbleibenden Lebensdauer konvektiver Zellen im Sinne des *Nowcastings* als sehr schwierig gestaltet, da deswegen die statistischen (linearen) Korrelationen zwischen den Eigenschaften detektierter Zellen und der Lebensdauer sehr gering sind. Einzelzellen, die meist eine Lebensdauer von insgesamt 30 – 60 min haben, können nach den Folgerungen jener Autoren nicht von Zellen mit einer längeren Lebensdauer unterschieden werden, wie z. B. von Superzellen. Besonders Zellen mit einer langen Lebensdauer und/oder einer großen Ausdehnung weisen jedoch das größte Schadenpotential auf.
Die Untersuchungen in der vorliegenden Arbeit zum mittleren zeitlichen Verlauf der Fläche konvektiver Zellen ordnen sich gut in das konzeptionelle Lebenszyklusmodell von Byers und Braham (1948) bzw. Doswell (1985) ein und bestätigen die Ergebnisse vergangener Studien (z. B. Weusthoff und Hauf, 2008; Wapler, 2021): Der Verlauf der Zellfläche kann gut durch eine nach unten geöffnete Parabelschar approximiert werden, die das anfängliche Wachstum der Zellen bis zu einer maximalen Fläche und die anschließende Dissipation widerspiegelt. Gleichzeitig weichen die einzelnen Lebenszyklen teils sehr deutlich von diesen mittleren Verläufen ab, was als Ausdruck der großen Variabilität interpretiert werden kann und dementsprechend eine schlechte Vorhersagbarkeit impliziert. Weitere Analysen

konnten zeigen, dass eine hohe Verlagerungsgeschwindigkeit der Zellen auf eine längere Lebensdauer hindeutet. Außerdem verfestigten diese Untersuchungen die Hinweise von Davini et al. (2012), dass gerade eine anfängliche, schnelle Vergrößerung der Zellfläche mit einer längeren Lebensdauer einhergeht.

Anhand der hier verwendeten Datensätze wurde der Einfluss der thermischen Stabilität der Atmosphäre auf eine solche schnelle Vergrößerung der Zellfläche gezeigt. Weitere Untersuchungen zum Potential der verschiedenen Umgebungsvariablen, zwischen Zellen mit eher kleinerer und größerer Zellfläche zu unterscheiden, unterstreichen die Bedeutung thermodynamischer Variablen für das Wachstum einer Zelle. Auch die vertikale Windscherung als ein wesentliches Element für die Zellorganisation spielt hierbei eine Rolle, wie sich auch in den Vergleichen der maximalen Zellfläche mit den kombinierten dynamisch-thermodynamischen Indizes SCP und SHIP zeigte. Besonders diese beiden Indizes, verschiedene Maße der vertikalen Windscherung und die Stärke der mittleren Grundströmung, mit der die oben erwähnte Verlagerungsgeschwindigkeit der Zellen assoziiert werden kann, unterscheiden sich für Zellen mit kurzer und langer Lebensdauer am deutlichsten. Quantitativ machen bivariate Analysen der mittleren Lebensdauer in Abhängigkeit von der thermischen Stabilität und der Windscherung allerdings deutlich, dass die Schärfe dieses kombinierten Unterscheidungsvermögens begrenzt ist. Bei hoher Instabilität und hoher Windscherung ist die Lebensdauer im Mittel nur etwa 10 – 15 min höher als bei weniger labilen und windschwachen Verhältnissen.

Motiviert durch die Ergebnisse dieser Analysen wurden multivariate Verfahren der Statistik und des maschinellen Lernens verwendet, um herauszufinden, ob durch eine Kombination der Informationen über die verschiedenen Zellattribute und die Umgebungsvariablen (Prädiktoren) besser zwischen Zellen mit kurzer und langer Lebensdauer oder kleinen und großen Zellflächen unterschieden werden kann. Um die große Anzahl der nachprozessierten Umgebungsvariablen zu reduzieren, wurden eine

Korrelations- und eine Clusteranalyse durchgeführt. Dadurch wurden zum einen lineare und rangbezogene Korrelationen zwischen je zwei Umgebungsvariablen bestimmt. Zum anderen konnten verschiedene Cluster von Umgebungsvariablen identifiziert werden, die eine Einteilung der Variablen in Gruppen ähnlicher Eigenschaften bezüglich einer Korrelationsoptimierung vornahmen. So konnten die bivariaten Korrelationen auf ein multivariates Bild erweitert werden.

Zusammen mit den uni- und bivariaten Analysen zum Unterscheidungsvermögen bildete die Clusteranalyse die Grundlage für die Auswahl der Umgebungsvariablen für die multivariaten Verfahren zur Abschätzung der Lebensdauer und der Zellfläche. Um eine differenzierte und robuste Einschätzung der Vorhersagbarkeit und der Bedeutung unterschiedlicher Prädiktoren für die Vorhersage zu erhalten, wurden im Rahmen der vorliegenden Arbeit das binäre Klassifikationsverfahren der multivariaten logistischen Regression, ein nicht-linearer Polynomansatz als Regressionsverfahren sowie der *Random Forest* nach Breiman (2001) angewendet. Für verschiedene Fragestellungen, die sowohl aus erkenntnistheoretischer Perspektive als auch für eine potentielle Anwendung der Ergebnisse zur Verbesserung von *Nowcasting*-Verfahren interessant sind, können verschiedene Verfahren unterschiedlich relevant sein. Während eine binäre Klassifikation beispielsweise darüber Aufschluss gibt, ob eine konvektive Zelle in Abhängigkeit von bestimmten Prädiktoren (ausgewählte Zellattribute und Umgebungsvariablen) eine eher kurze oder lange Lebensdauer zu erwarten hat (mit einem entsprechenden Trennwert zwischen kurz und lang), verknüpft eine Regression die Prädiktoren mit einem kontinuierlichen Wert für die Lebensdauer. Der *Random Forest* ist so konzipiert, dass er sowohl eine binäre oder multikategorielle Klassifikation als auch eine Regression ermöglicht. Dadurch ist er für eine Reihe von Untersuchungen geeignet und konnte hier in beiden Modi angewendet und mit der binären logistischen Regression bzw. dem nicht-linearen Polynomansatz verglichen werden.

Um einen Schritt weiter als eine reine multivariate Analyse der Zusammenhänge zwischen den Prädiktoren und der Lebensdauer bzw. der maximalen Zellfläche zu gehen, wurde ein Verfahren entwickelt, welches die Generierung eines ganzen Ensembles bestehend aus vielen Modellen der logistischen Regression, des *Random Forests* oder des Polynomansatzes auf der Basis des kombinierten Datensatzes der knapp 40 000 Lebenszyklen und Umgebungsvariablen ermöglichte. Dazu wurde dieser jeweils in unterschiedliche Trainings- und Testdatensätze aufgespalten. Nach einer geeigneten Vorbereitung der Daten konnte damit nicht nur jeweils ein Modellensemble aufgestellt, sondern dieses im gleichen Zug mit unabhängigen Daten auf sein Vorhersagevermögen untersucht und evaluiert werden. Damit ließ sich beurteilen, wie gut ein solches Ensemble zur Abschätzung der Lebensdauer oder der maximalen Fläche konvektiver Zellen während ihres Lebenszyklus geeignet ist. Zudem spiegelten die Ergebnisse daher die zu erwartende Vorhersagegüte wider, die ein solches Verfahren in einer *Nowcasting*-Anwendung erreichen würde. Dieser Schritt ist für eine Studie besonders, die den Lebenszyklus konvektiver Zellen oder den Einfluss von Umgebungsbedingungen auf die Eigenschaften konvektiver Zellen analysiert. Ein solcher Ensembleansatz ermöglicht darüber hinaus probabilistische Vorhersagen, die eine Quantifizierung der Vorhersageunsicherheit inkludieren. Zudem erlauben die logistische Regression, der *Random Forest* und der Polynomansatz die Quantifizierung der Wichtigkeit der einzelnen Prädiktoren, deren Variabilität durch den Ensembleansatz ebenfalls erfasst werden konnte.

Ein Vergleich der binären Klassifikation zwischen dem *Random Forest* und der logistischen Regression zeigte, dass beide Verfahren ähnlich gute Vorhersagen treffen, die logistische Regression im deterministischen Sinn leicht bessere Werte für die Gütemaße erhält. Da die logistische Regression auf einer globalen Optimierung der Modellparameter beruht, während der *Random Forest* aufgrund seines lokalen Optimierungsansatzes den Raum der Prädiktoren

in feinere Unterräume unterteilen kann, erreicht letzterer daher bessere Werte für probabilistische Gütemaße. Mit beiden Verfahren wurden deutlich bessere Vorhersagen erzielt als bei einer zufälligen oder einer Persistenzvorhersage. Dies gilt sowohl für die Abschätzung der Lebensdauerklasse (kurz/lang mit Trennwert 60 min) als auch der Klasse der maximalen Zellfläche (klein/groß mit Trennwert 60 km^2). Die maximale Zellfläche ließ sich dabei besser abschätzen als die Lebensdauer der Zellen. Die Gütemaße sind jedoch inhärent abhängig von der Wahl des Trennwerts zwischen den Klassen sowie dem genauen Vorgehen bei der Evaluation und müssen daher mit Sorgfalt interpretiert werden (vgl. Kapitel 6). Besonders gut ist die Klassifikation der Lebensdauer bei solchen Modellensembles, in denen die Information über die Zellfläche und die Fläche des Zellkerns zu Beginn des Lebenszyklus als Prädiktor in die Modellbildung eingeht. Dies bestätigt das Indiz aus den oben geschilderten Lebenszyklusanalysen, dass ein schnelles anfängliches Wachstum einer Zelle zu einer längeren Lebensdauer führt. Als ähnlich wichtig für die Klassifikation der Lebensdauer wie diese beiden Zellattribute stellten sich sowohl für die logistische Regression als auch den *Random Forest* erwartungsgemäß solche Umgebungsvariablen heraus, welche die vertikale Windscherung oder die mittlere Strömung beschreiben. Für die Klassifikation der maximalen Zellfläche ist der Einfluss aller Umgebungsvariablen auf die Vorhersagegüte vernachlässigbar. Hier deutet die Evaluation darauf hin, dass bereits allein aus dem anfänglichen Wachstum abgeleitet werden kann, ob eine Zelle in ihrer weiteren Entwicklung eine große Fläche erreichen wird. Qualitativ ähnliche Rückschlüsse bezüglich der Wichtigkeit der einzelnen Prädiktoren lassen sich aus dem Vergleich der Ensemblevorhersagen für die Regressionsverfahren ziehen. Ebenso lässt sich mit diesen Verfahren die maximale Zellfläche insgesamt besser abschätzen als die Lebensdauer der Zellen. Der *Random Forest* weist für beide Prädiktanden insgesamt leicht bessere Vorhersagen hinsichtlich der Auflösung, Verlässlichkeit und des Vorhersage-Bias auf als ein linearer oder nicht-linearer Polynomansatz

fünfter Ordnung. Die Abschätzung der Lebensdauer zeigt allerdings mit einem sogenannten einfachen Strömungsfeldansatz (vgl. Kapitel 5.1.2), der keine Umgebungsvariablen berücksichtigt, einen geringeren Vorhersage-Bias als mit dem Polynomansatz oder dem *Random Forest*. Aufgrund seiner Konstruktion beträgt jedoch mit diesem Ansatz die höchste Lebensdauer-Vorhersage weniger als anderthalb Stunden, was deutlich unter der beobachteten Lebensdauer der langlebigsten Zellen im Datensatz liegt. Mit dem *Random Forest* werden als höchste Lebensdauer-Vorhersage maximal etwas über zwei Stunden und mit dem Polynomansatz mehr als drei Stunden erreicht. Für die beiden letzten Verfahren kann der Wertebereich der Vorhersagen zudem über ein geeignetes *Resampling* gesteuert werden. Die Schärfe der Vorhersagen ließ sich im Vergleich zu den Unterschieden der mittleren Lebensdauer in den oben beschriebenen bivariaten Analysen durch die Anwendung der multivariaten Verfahren bzw. des Strömungsfeldansatzes in etwa verdoppeln, liegt mit etwa 20 – 25 min jedoch weiterhin recht niedrig. Die maximale Zellfläche weist eine hohe Korrelation zu der Zellfläche zum Zeitpunkt 15 min nach der ersten Detektion auf. Das Ergebnis aus den Klassifikationsverfahren lässt sich somit darauf erweitern, dass alleine aus dem anfänglichen Wachstum gut abgeleitet werden kann, auf welche maximale Fläche eine Zelle anwächst. Es konnte jedoch auch gezeigt werden, dass die Berücksichtigung von Umgebungsvariablen die ohnehin schon hohe Vorhersagegüte noch weiter erhöhen kann, insbesondere zu Beginn der Zellentwicklung. Im linearen Polynomansatz zeigen dabei Maße der thermischen Instabilität bzw. der für Konvektion verfügbaren Energie einen bedeutenden Einfluss, was mit dem oben analysierten schnelleren Wachstum der Zellen bei höherer thermischer Instabilität übereinstimmt.

Im Sinne einer quantitativen Vorhersage der Lebensdauer ist das Vorhersagevermögen der Modellensembles rein basierend auf bestimmten Umgebungsvariablen insgesamt eher begrenzt. Die Ergebnisse der Studie von MacKeen et al. (1999), die zeigen, dass die Zusammenhänge zwischen der Lebensdauer konvektiver Zellen und deren Zelleigenschaften zu gering

sind, um eine Unterscheidung zwischen kurz- und langlebigen Zellen zu treffen, werden durch die vorliegende Arbeit auf einen aktuelleren und differenzierteren Stand gebracht. Beispielsweise kann rein auf der Basis von Umgebungsvariablen mit einer Wahrscheinlichkeit von über 60 % korrekt abgeschätzt werden, ob eine Zelle mindestens 60 min lebt oder nicht. Rund 15 min nach der ersten Detektion einer Zelle kann unter Hinzunahme der Information über die Entwicklung der Zellfläche sogar mit einer Wahrscheinlichkeit von über 70 % korrekt abgeschätzt werden, ob diese Zelle noch mindestens weitere 45 min lebt oder nicht. Nimmt man in Kauf, dass Zellen mit einer kurzen Lebensdauer zu 50 % fälschlicherweise als langlebige Zellen vorhersagt werden, ist es möglich, 80 bis 90 % aller Zellen mit einer langen Lebensdauer bereits nach 15 min zu identifizieren. Mit dem Wissen, dass insgesamt viel mehr Zellen mit einer kurzen als mit einer langen Lebensdauer auftreten, bedeuten diese Ergebnisse jedoch weiterhin ein relativ großes Verhältnis von Fehlalarmen zu korrekten Zuordnungen.

Die ausführliche Evaluation der verschiedenen Modellensembles hatte zum Ziel, neben dem wissenschaftlichen Erkenntnisgewinn einen differenzierten Blick auf die Potentiale und Grenzen zu ermöglichen, welche für die Verbesserung von *Nowcasting*-Verfahren im Sinne einer Verbesserung der Abschätzung der Lebensdauer und der maximalen Fläche konvektiver Zellen bestehen. Diese Abschätzungen können zum einen insbesondere für die Verbesserung der internen Lebenszyklusmodelle von Zellverfolgungsalgorithmen von Interesse sein, sind aber auch direkt für die genauere zeitliche und räumliche Spezifizierung von Gewitterwarnungen von großer Relevanz. Die vorliegende Arbeit verdeutlicht damit einige mögliche Vorteile eines integrierten Vorhersagesystems für den Bereich der Kürzestfristvorhersage, in dem nahtlose Vorhersagen der konvektiven Aktivität auf der Basis einer Kombination aus räumlich und zeitlich hochaufgelösten Vorhersagen von numerischen Wettervorhersagemodellen und *Nowcasting*-Verfahren möglich sind. In diesem Zusammenhang liefert die vorliegende Arbeit

einen wichtigen Beitrag im Rahmen der Optimierung von automatisierten Warnprozessen, wie sie durch viele Wetterdienste operationell betrieben werden. Entscheidungsträger, Unternehmen und Privatpersonen profitieren gleichermaßen von möglichst präzisen Warnungen vor den gefährlichen Begleiterscheinungen konvektiver Zellen.

Die gezeigten Analysen und Modellstudien stellen nur einen Teil der im Rahmen dieser Arbeit durchgeführten Untersuchungen dar. Beispielsweise wurden auch Modellensembles auf der Basis des *Random Forests* untersucht, welche eine Klassifikation der Lebensdauer in mehr als nur zwei Klassen vornehmen. Die Evaluation solcher Realisierungen mit drei, vier und fünf Klassen deutet auf ein Vorhersagepotential hin, welches für spezifische Anwendungen von Interesse sein könnte. Darüber hinaus kann die Vorhersagegüte der verwendeten Verfahren noch gesteigert werden, indem für jedes dieser Verfahren eine systematische Auswahl der Prädiktoren durch bekannte Vorgehensweisen wie die schrittweise Regression durchgeführt wird. Ein weiteres Potential zur Erhöhung der Vorhersagegüte könnte in der Berücksichtigung der räumlichen Verteilung der Umgebungsvariablen in Bezug auf die Position der konvektiven Zellen liegen, welche unter anderem Sherburn et al. (2016) und Kunz et al. (2020) analysierten. Hierzu wäre die Anwendung von anderen Verfahren aus dem Bereich der Statistik oder des maschinellen Lernens erforderlich, wie beispielsweise von konvolutionalen neuronalen Netzen.
Weiterentwickelte Verfahren der Zellverfolgung wie z. B. KONRAD3D (DWD) werden ferner in der Lage sein, die Lebenszyklen konvektiver Zellen realistischer und mit einer größeren Vielfalt von weiteren Zellattributen zu beschreiben. Neben der Information über die Vertikalstruktur und den Flüssigwassergehalt der Zellen können auch die mit den modernen *Dual-Pol* Doppler-Radargeräten gewonnenen Informationen zur Hydrometeorklassifikation genutzt werden. Dies eröffnet einen großen Raum weiterer Möglichkeiten für die statistische Lebenszyklusanalyse und -vorhersage, sobald Daten über einen genügend großen Stichprobenzeitraum

zur Verfügung stehen. Darüber hinaus stellen Verfahren zur Abschätzung verschiedener Zellattribute für das *Nowcasting* konvektiver Zellen auf der Basis von simulierten Zellen aus den NWV-Vorhersagen einen weiteren interessanten Ansatz dar (vgl. Feige et al., 2018; DWD, 2021c). Im Sinne des Multi-Daten-Ansatzes erscheinen außerdem kombinierte Analysen und Verfahren als vielversprechend, die auf der Basis von Satelliten-, Radar-, und Blitzdaten und/oder Daten aus *Nowcasting*-Verfahren und numerischen Wettervorhersagemodellen ein multidimensionales Bild der messbaren Eigenschaften von konvektiven Zellen zeichnen (z. B. Nisi et al., 2014; Zöbisch et al., 2020).

Akronymverzeichnis

Akronym	Bedeutung	Seite
ACC	Balanciertes Maß für die Genauigkeit (*Accuracy*)	246
AUC	Fläche unterhalb der ROC-Kurve (*Area Under the Curve*)	130
B	Bias	128
BRMSE	Balancierter *RMSE* (*Balanced Root Mean Squared Error*)	248
BS	Brier Score	133
BSS	Brier Skill Score	133
CSI	*Critical Success Index*	128
DAT	Datums- und Uhrzeitangabe für ein Zellobjekt	159
DIS	Unterscheidungsfaktor (*Discrimination Factor*)	217
ETS	*Equitable Threat Score*	130
F	Fehlalarmrate (*False Alarm Rate*)	128
FAR	Fehlalarmverhältnis (*False Alarm Ratio*)	128
GI	Gini Index	110
H	Trefferrate (*Hit Rate*)	128
HSS	Heidke Skill Score	130
ID_Z	KONRAD-interne Identifikationsnummer eines Zellobjekts	158
KE	Kreuz-Entropie	110
MF	Missklassifikationsfehler	110
MSE	Mittlerer quadratischer Fehler (*Mean Squared Error*)	95
OLP	Überlappung (*Overlap*)	217
OR	Chancenverhältnis (*Odds Ratio*)	128

PC	*Proportion Correct* 128
$RMSE$	Wurzel aus dem mittleren quadratischen Fehler (*Root Mean Squared Error*) 95
RTM	Indikator, welche Radardaten 5 min vor dem Detektionszeitpunkt nicht verfügbar sind 159
RTN	Indikator, welche Radardaten zum jeweiligen Detektionszeitpunkt nicht verfügbar sind 159
RTP	Indikator, welche Radardaten 5 min nach dem Detektionszeitpunkt nicht verfügbar sind 159
SR	Erfolgsverhältnis (*Success Ratio*) 128
TSS	*True Skill Statistic* 128
BRN	*Bulk Richardson Number* 45
CAPE	Konvektiv verfügbare potentielle Energie (*Convective Available Potential Energy)* 36
CIN	Konvektive Hemmung (*Convective Inhibition*) 36
COSMO	*Consortium for Small-Scale Modelling* 149
DCI	*Deep Convective Index* 176
DLS	Betrag der vertikalen Windvektordifferenz (0 – 6 km) (*Deep Layer Shear*) 38
DWD	Deutscher Wetterdienst 9
ESWD	*European Severe Weather Database* 4
HKN	Hebungskondensationsniveau 30
IWV	Vertikal integrierter Wasserdampfgehalt (*Integrated Water Vapor*) 73
K15	Modellstudie mit einer Kombination der Prädiktoren von Z15 und U2 (LOGR) bzw. U15 (POLY, RF) 272
$K15^{+}$	Modellstudie mit einer Kombination der Prädiktoren von $Z15^{+}$ und U2 (LOGR) bzw. U15 (POLY, RF) 272
$K15^{+}_{var}$	Modellstudie mit denselben Prädiktoren wie $K15^{+}$, nur mit Variation des Entscheidungstrennwerts μ 272
K5	Modellstudie mit einer Kombination der Prädiktoren von Z5 und U2 (LOGR) bzw. U15 (POLY, RF) 272

KKN	Cumulus-Kondensationsniveau	30
KONRAD	Zellverfolgungsalgorithmus (Konvektionsentwicklung in Radarprodukten)	10
LI	*Lifted Index*	72
LLS	Betrag der vertikalen Windvektordifferenz (0 – 1 km) (*Low Level Shear*)	175
LOGR	Logistische Regression	252
LR	Mittlerer vertikaler Temperaturgradient (einer Schicht; *Lapse Rate*)	71
MCC	Mesoskaliger Konvektiver Komplex (*Mesoscale Convective Complex*)	65
MCS	Mesoskaliges Konvektives System (*Mesoscale Convective System*)	6
ML	Mischungsschicht (*Mixed Layer*)	30
MLS	Betrag der vertikalen Windvektordifferenz (0 – 3 km) (*Medium Layer Shear)*	175
MU	*Most Unstable*	30
NFK	Niveau freier Konvektion	30
NNA	Niveau des neutralen Aufstiegs	32
NWV	Numerische Wettervorhersage	7
OOB	*Out of Bag*	116
OSP	*Oversampling*	120
PA	Parabelansatz	287
PAM	*Partitioning Around Medoids*	91
POLY	Polynomansatz	279
RF	*Random Forest*	252
ROC	*Relative/Receiver Operating Characteristic*	130
RSP	*Resampling*	119
SCP	*Supercell Composite Parameter*	73
SF	Strömungsfeldansatz	289
SHIP	*Significant Hail Parameter*	74

SI	*Showalter Index*	174
SLI	*Surface Lifted Index*	72
SMOTE	*Synthetic Minority Oversampling Technique*	123
SRH	Sturm-relative Helizität	62
TT	*Total Totals*	176
U15	Modellstudie mit 15 Umgebungsvariablen	271
U2	Zweite Modellstudie mit zwei Umgebungsvariablen	271
U2_0	Erste Modellstudie mit zwei Umgebungsvariablen	250
U6	Modellstudie mit sechs Umgebungsvariablen	271
USP	*Undersampling*	120
VT	*Vertical Totals*	176
Z15	Modellstudie mit der Zellfläche und der Fläche des Zellkerns zum Zeitpunkt 15 min nach der ersten Detektion im Radarbild (LOGR, POLY, RF) bzw. nur mit der Zellfläche (PA, SF)	271
$Z15^{+}$	Modellstudie mit der Zellfläche, der Fläche des Zellkerns und der Verlagerung des Zellobjekts nach 15 min	271
Z5	Modellstudie mit der Zellfläche und der Fläche des Zellkerns zum Zeitpunkt 5 min nach der ersten Detektion im Radarbild (LOGR, POLY, RF) bzw. nur mit der Zellfläche (PA, SF)	271

Literaturverzeichnis

Achatz, U. und G. Schmitz, 1997: On the closure problem in the reduction of complex atmospheric models by PIPs and EOFs: A comparison for the case of a two-layer model with zonally symmetric forcing. *Journal of the Atmospheric Sciences*, **54**, 2452–2474.

Andersson, T. M., C. J. Andersson, C. Jacobsson, und S. Nilsson, 1989: Thermodynamic indices for forecasting thunderstorms in southern Sweden. *Meteorological Magazine*, **118**, 141–146.

Arakawa, C. und V. R. Lamb, 1977: Computational Design of the Basic Dynamical Processes of the UCLA General Circulation Model. *General Circulation Models of the Atmosphere*, J. Chang, Hrsg., Elsevier, Amsterdam, Niederlande, S. 173–265.

Atanassov, K., 2014: *Index Matrices: Towards an Augmented Matrix Calculus*. Springer International Publishing, Basel, Schweiz, 110 S.

Atkins, N. T. und R. M. Wakimoto, 1991: Wet microburst activity over the southeastern United States: Implications for forecasting. *Weather and Forecasting*, **6**, 470–482.

Backhaus, K., B. Erichson, W. Plinke, und R. Weiber, 2016: *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, 647 S.

Backhaus, K., B. Erichson, und R. Weiber, 2015: *Fortgeschrittene Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, 454 S.

Baldauf, M., J. Förstner, S. Klink, T. Reinhardt, C. Schraff, A. Seifert, und K. Stephan, 2016: Kurze Beschreibung des Lokal-Modells Kürzestfrist COSMO-DE (LMK) und seiner Datenbanken auf dem Datenserver des DWD, Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/SharedDocs/downloads/DE/`

`modelldokumentationen/nwv/cosmo_de/cosmo_de_dbbeschr_version_2_4_161124.pdf?__blob=publicationFile&v=4`, abgerufen: 07. Januar 2021.

Barlow, W., 1993: A New Index for the Prediction of Deep Convection. *Preprints, 17th Conference on Severe Local Storms (St. Louis, USA)*, American Meteorological Society, Boston, USA, S. 129–132.

Bartels, H., E. Weigl, T. Reich, P. Lang, A. Wagner, O. Kohler, und N. Gerlach, 2004: Projekt RADOLAN – Routineverfahren zur Online-Aneichung der Radarniederschlagsdaten mit Hilfe von automatischen Bodenniederschlagsstationen (Ombrometer), Abteilung Hydrometeorologie, Deutscher Wetterdienst, Offenbach, Deutschland.

Barthlott, C., B. Mühr, und C. Hoose, 2017: Sensitivity of the 2014 Pentecost storms over Germany to different model grids and microphysics schemes. *Quarterly Journal of the Royal Meteorological Society*, **143**, 1485–1503.

Batista, G., R. Prati, und M. Monard, 2004: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, **6**, 20–29.

Bechtold, P., E. Bazile, F. Guichard, P. Mascart, und E. Richard, 2001: A mass-flux convection scheme for regional and global models. *Quarterly Journal of the Royal Meteorological Society*, **127**, 869–886.

Bedka, K. M., 2011: Overshooting cloud top detections using MSG SEVIRI infrared brightness temperatures and their relationship to severe weather over Europe. *Atmospheric Research*, **99**, 175–189.

Bernard, S., L. Heutte, und S. Adam, 2009: Influence of Hyperparameters on Random Forest Accuracy. *Multiple Classifier Systems*, J. A. Benediktsson, J. Kittler, und F. Roli, Hrsg., Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, S. 171–180.

Bjerknes, J., 1938: Saturated-adiabatic ascent of air through dry-adiabatically descending environment. *Quarterly Journal of the Royal Meteorological Society*, **64**, 325–330.

Blahak, U., 2005: Analyse des Extinktionseffektes bei Niederschlagsmessungen mit einem C-Band Radar anhand von Simulation und Messung. Dissertation, Fakultät für Physik, Universität Karlsruhe, Karlsruhe, Deutschland.

Blahak, U. und A. de Lozar, 2020: EMVORADO – Efficient Modular VOlume scan RADar Operator: A user's guide. Consortium for Small-Scale Modelling. URL: `http://www.cosmo-model.org/content/model/documentation/core/emvorado_userguide.pdf`, abgerufen: 07. Januar 2021.

Blahak, U., et al., 2018: Development of a new seamless prediction system for very short range convective-scale forecasting at DWD. Geophysical Research Abstracts. Vol. 20, EGU2018-9642, Wien, Österreich.

Bolton, D., 1980: The computation of equivalent potential temperature. *Monthly Weather Review*, **108**, 1046–1053.

Box, G. E. P. und D. R. Cox, 1964: An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252.

Breiman, L., 1996: Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L., 2000: Randomizing outputs to increase prediction accuracy. *Machine Learning*, **40**, 229–242.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5–32.

Breiman, L., J. Friedman, R. Olshen, und C. Stone, 1984: *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, USA, 357 S.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Brisson, E., C. Brendel, S. Herzog, und B. Ahrens, 2018: Lagrangian evaluation of convective shower characteristics in a convection-permitting model. *Meteorologische Zeitschrift*, **27**, 59–66.

Bronstert, A., A. Agarwal, B. Bossenkool, M. Peter, M. Heistermann, L. Köhn-Reich, T. Moran, und D. Wendi, 2017: Die Sturzflut von Braunsbach am 29. Mai 2016 – Entstehung, Ablauf und Schäden eines „Jahrhundertereignisses“. Teil 1: Meteorologische und hydrologische Analyse. *Hydrologie und Wasserbewirtschaftung*, **61**, 150–162.

Brooks, H. E., 2007: Ingredients-Based Forecasting. *Atmospheric Convection: Research and Operational Forecasting Aspects*, D. B. Giaiotti, R. Steinacker, und F. Stel, Hrsg., Springer-Verlag Wien, Wien, Österreich, S. 133–140.

Brooks, H. E., 2009: Proximity soundings for severe convection for Europe and the United States from reanalysis data. *Atmospheric Research*, **93**, 546–553.

Brooks, H. E., A. R. Anderson, K. Riemann, I. Ebbers, und H. Flachs, 2007: Climatological aspects of convective parameters from the NCAR/NCEP reanalysis. *Atmospheric Research*, **83**, 294–305.

Brooks, H. E., J. W. Lee, und J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, **67–68**, 73–94.

Browning, K. A., 1965: *A Family Outbreak of Severe Local Storms: a Comprehensive Study of the Storms in Oklahoma on 26 May 1963*, Vol. 65. Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force, Dayton, USA, 346 S.

Browning, K. A., 1977: The Structure and Mechanisms of Hailstorms. *Hail: A Review of Hail Science and Hail Suppression*, G. B. Foote und C. A. Knight, Hrsg., American Meteorological Society, Boston, USA, S. 1–47.

Bryan, G. H., J. C. Knievel, und M. D. Parker, 2006: A multimodel assessment of RKW theory's relevance to squall-line characteristics. *Monthly Weather Review*, **134**, 2772–2792.

Bunkers, M. J., 2002: Vertical wind shear associated with left-moving supercells. *Weather and Forecasting*, **17**, 845–855.

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, und M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. *Weather and Forecasting*, **15**, 61–79.

Burgess, D. W. und E. B. Curran, 1985: The Relationship of Storm Type to Environment in Oklahoma on 26 April 1984. *Preprints, 14th Conference on Severe Local Storms (Indianapolis, USA)*, American Meteorological Society, Boston, USA, S. 208–211.

Byers, H. R. und R. R. J. Braham, 1948: Thunderstorm structure and circulation. *Journal of Meteorology*, **5**, 71–86.

Cao, Z., 2008: Severe hail frequency over Ontario, Canada: Recent trend and variability. *Geophysical Research Letters*, **35**, L14 803.

Carlson, T. N., S. G. Benjamin, G. S. Forbes, und Y. F. Li, 1983: Elevated mixed layers in the regional severe storm environment: Conceptual model and case studies. *Monthly Weather Review*, **111**, 1453–1474.

Charba, J. P., 1977: *Operational system for predicting thunderstorms two to six hours in advance*. Techniques Development Laboratory, National Weather Service, Silver Spring, USA.

Chawla, N. V., K. W. Bowyer, L. O. Hall, und W. P. Kegelmeyer, 2002: SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Chronis, T., L. D. Carey, C. J. Schultz, E. V. Schultz, K. M. Calhoun, und S. J. Goodman, 2015: Exploring lightning jump characteristics. *Weather and Forecasting*, **30**, 23–37.

Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Craven, J. P., R. E. Jewell, und H. E. Brooks, 2002: Comparison between observed convective cloud-base heights and lifting condensation level for two different lifted parcels. *Weather and Forecasting*, **17**, 885–890.

Cressman, G. P., 1959: An operational objective analysis system. *Monthly Weather Review*, **87**, 367–374.

Czernecki, B., M. Taszarek, M. Marosz, M. Półrolniczak, L. Kolendowicz, A. Wyszogrodzki, und J. Szturc, 2019: Application of machine learning to large hail prediction – The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmospheric Research*, **227**, 249–262.

Davies, H. C. und R. E. Turner, 1977: Updating prediction models by dynamical relaxation: an examination of the technique. *Quarterly Journal of the Royal Meteorological Society*, **103**, 225–245.

Davies-Jones, R., 1984: Streamwise vorticity: The origin of updraft rotation in supercell storms. *Journal of Atmospheric Sciences*, **41**, 2991–3006.

Davini, P., R. Bechini, R. Cremonini, und C. Cassardo, 2012: Radar-based analysis of convective storms over northwestern Italy. *Atmosphere*, **3**, 33–58.

Davis, C. A. und S. B. Trier, 2007: Mesoscale convective vortices observed during BAMEX. Part I: Kinematic and thermodynamic structure. *Monthly Weather Review*, **135**, 2029–2049.

Dixon, M. und G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis, and nowcasting – A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, **10**, 785–797.

Doms, G. und M. Baldauf, 2018: A description of the nonhydrostatic regional COSMO-Model – Part I: Dynamics and numerics. Consortium for Small-Scale Modelling. URL: `http://www.cosmo-model.org/content/model/documentation/core/cosmo_dynamics_5.05.pdf`, abgerufen: 07. Januar 2021.

Doms, G., et al., 2018: A description of the nonhydrostatic regional COSMO-Model – Part II: Physical parameterizations. Consortium for Small-Scale Modelling. URL: `http://www.cosmo-model.org/content/model/documentation/core/cosmo_physics_5.05.pdf`, abgerufen: 07. Januar 2021.

Doswell, C. A., 1985: *The Operational Meteorology of Convective Weather. Storm Scale Analysis. Vol. 2.* Air Weather Service, Scott Air Force Base, Belleville, USA.

Doswell, C. A., 1987: The distinction between large-scale and mesoscale contribution to severe convection: A case study example. *Weather and Forecasting*, **2**, 3–16.

Doswell, C. A., 2007: Historical overview of severe convective storms research. *E-Journal of Severe Storms Meteorology*, **2**, 1–25.

Doswell, C. A., H. E. Brooks, und R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11**, 560–581.

Doswell, C. A., R. Davies-Jones, und D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, **5**, 576–585.

Doswell, C. A. und P. M. Markowski, 2004: Is buoyancy a relative quantity? *Monthly Weather Review*, **132**, 853–863.

Doswell, C. A. und E. N. Rasmussen, 1994: The effect of neglecting the virtual temperature correction on CAPE calculations. *Weather and Forecasting*, **9**, 625–629.

Dotzek, N., P. Groenemeijer, B. Feuerstein, und A. M. Holzer, 2009: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD. *Atmospheric Research*, **93**, 575–586.

Draper, N. R. und H. Smith, 1998: *Applied Regression Analysis*. Wiley Series in Probability and Statistics, John Wiley & Sons, New York, USA, 706 S.

Droegemeier, K. K., S. M. Lazarus, und R. Davies-Jones, 1993: The influence of helicity on numerically simulated convective storms. *Monthly Weather Review*, **121**, 2005–2029.

DWD, 2020: RADOLAN/RADVOR: Hoch aufgelöste Niederschlagsanalyse und -vorhersage auf der Basis quantitativer Radar- und Ombrometerdaten für grenzüberschreitende Fluss-Einzugsgebiete von Deutschland im Echtzeitbetrieb – Beschreibung des Kompositformats, Version 2.5.1, Abteilung Hydrometeorologie, Deutscher Wetterdienst, Offenbach, Deutschland.

DWD, 2021a: Assimilation von Radar-Niederschlagdaten im COSMO-EU. Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/DE/fachnutzer/forschung_lehre/numerische_wettervorhersage/nwv_aenderungen/_functions/DownloadBox_modellaenderungen/cosmo_eu/pdf_2011_2015/pdf_lme_03_09_2014.pdf?__blob=publicationFile&v=7`, abgerufen: 07. Januar 2021.

DWD, 2021b: Der DWD/Messnetz/Atmosphärenbeobachtung/Radarverbund. Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/DE/derdwd/messnetz/atmosphaerenbeobachtung/_functions/Teasergroup/radarverbund_teaser5.html`, abgerufen: 07. Januar 2021.

DWD, 2021c: Forschung/DWD-Forschungsprogramme/„SINFONY“. Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/DE/forschung/forschungsprogramme/sinfony_iafe/sinfony_node.html`, abgerufen: 07. Januar 2021.

DWD, 2021d: Forschung/Wettervorhersage/Meteorologische Fachverfahren/Nowcasting-Verfahren/Konvektive Entwicklung (KONRAD). Deutscher Wetterdienst, Offenbach, Deutschland. URL: https://www.dwd.de/DE/forschung/wettervorhersage/met_fachverfahren/nowcasting/konrad_node.html, abgerufen: 07. Januar 2021.

DWD, 2021e: Forschung/Wettervorhersage/Meteorologische Fachverfahren/Unterstützung des Warnprozesses mit AutoWARN. Deutscher Wetterdienst, Offenbach, Deutschland. URL: https://www.dwd.de/DE/forschung/wettervorhersage/met_fachverfahren/unterstuetzung_warnprozess_autowarn/unterstuetzung_warnprozess_autowarn_node.html, abgerufen: 07. Januar 2021.

Feger, R., M. Werner, R. Posada, K. Wapler, und U. Blahak, 2019: Generation of an object-based nowcasting ensemble. 3rd European Nowcasting Conference, Madrid, Spanien. URL: https://www.eumetnet.eu/european-nowcasting-conference-enc-2019/, abgerufen: 07. Januar 2021.

Feige, K., R. Posada, und U. Blahak, 2018: Developing a Concept to Visualize Object-based Weather Forecasting Ensembles. *Workshop on Visualisation in Environmental Sciences (EnvirVis, Brno, Tschechische Republik)*, K. Rink, D. Zeckzer, R. Bujack, und S. Jänicke, Hrsg., The Eurographics Association, Genf, Schweiz, S. 19–25.

Feng, Z., X. Dong, B. Xi, S. A. McFarlane, A. Kennedy, B. Lin, und P. Minnis, 2012: Life cycle of midlatitude deep convective systems in a Lagrangian framework. *Journal of Geophysical Research: Atmospheres*, **117**, D23.

Fink, A. H., T. Brücher, V. Ermert, A. Krüger, und J. G. Pinto, 2009: The European storm Kyrill in January 2007: Synoptic evolution, meteorological impacts and some considerations with respect to climate change. *Natural Hazards and Earth System Sciences*, **9**, 405–423.

Finley, J. P., 1884: Tornado prediction. *American Meteorological Journal*, **1**, 85–88.

Fluck, E., 2018: Hail statistics for European countries. Dissertation, Fakultät für Physik, Karlsruher Institut für Technologie (KIT), Karlsruhe, Deutschland.

Fovell, R. G. und P. S. Dailey, 1995: The temporal behavior of numerically simulated multicell-type storms. Part I. Modes of behavior. *Journal of Atmospheric Sciences*, **52**, 2073–2095.

Fovell, R. G. und P.-H. Tan, 1998: The temporal behavior of numerically simulated multicell-type storms. Part II: The convective cell life cycle and cell regeneration. *Monthly Weather Review*, **126**, 551–577.

Fujita, T. T., 1978: *Manual of downburst identification for Project NIMROD*. Chicago University, Chicago, USA.

Gal-Chen, T. und R. C. J. Somerville, 1975: On the use of a coordinate transformation for the solution of the Navier-Stokes equations. *Journal of Computational Physics*, **17**, 209–228.

Galway, J. G., 1956: The Lifted Index as a predictor of latent instability. *Bulletin of the American Meteorological Society*, **37**, 528–529.

Gatzen, C. P., A. H. Fink, D. M. Schultz, und J. G. Pinto, 2020: An 18–year climatology of derechos in Germany. *Natural Hazards and Earth System Sciences*, **20**, 1335–1351.

Gensini, V. A. und M. K. Tippett, 2019: Global Ensemble Forecast System (GEFS) predictions of days 1–15 U.S. tornado and hail frequencies. *Geophysical Research Letters*, **46**, 2922–2930.

George, J. J., 1960: *Weather Forecasting for Aeronautics*. Academic Press, New York, USA, 673 S.

Gilbert, G. K., 1884: Finley's tornado predictions. *American Meteorological Journal*, **1**, 166–172.

Glahn, H. R., 1985: Statistical Weather Forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy und R. W. Katz, Hrsg., Westview Press, Boulder, USA, S. 289–335.

Greene, D. R. und R. A. Clark, 1972: Vertically integrated liquid water? A new analysis tool. *Monthly Weather Review*, **100**, 548–552.

Groenemeijer, P., 2009: Convective storm development in contrasting thermodynamic and kinematic environments. Dissertation, Fakultät für Physik, Universität Karlsruhe, Karlsruhe, Deutschland.

Groenemeijer, P., et al., 2017: Severe convective storms in Europe: Ten years of research and education at the European Severe Storms Laboratory. *Bulletin of the American Meteorological Society*, **98**, 2641–2651.

Groenemeijer, P. H. und A. van Delden, 2007: Sounding-derived parameters associated with large hail and tornadoes in the Netherlands. *Atmospheric Research*, **83**, 473–487.

Haklander, A. J. und A. van Delden, 2003: Thunderstorm predictors and their forecast skill for the Netherlands. *Atmospheric Research*, **67–68**, 273–299.

Hamann, U., et al., 2019: Nowcasting of thunderstorm severity with Machine Learning in the Alpine Region. 3rd European Nowcasting Conference, Madrid, Spanien. URL: `https://repositorio.aemet.es/handle/20.500.11765/10617`, abgerufen: 07. Januar 2021.

Hamilton, R. E., 1970: Use of Detailed Intensity Radar Data in Mesoscale Surface Analysis of the 4 July 1969 Storm in Ohio. *Preprints, 14th Conference on Radar Meteorology (Tucson, USA)*, American Meteorological Society, Boston, USA, S. 339–342.

Hand, W. H. und B. J. Conway, 1995: An object-oriented approach to nowcasting showers. *Weather and Forecasting*, **10**, 327–341.

Handwerker, J., 2002: Cell tracking with TRACE3D – A new algorithm. *Atmospheric Research*, **61**, 15–34.

Hansen, P., 2010: *Discrete Inverse Problems: Insight and Algorithms*. Fundamentals of Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, USA, 206 S.

Harrell, F. E. J., 2015: *Regression Modeling Strategies*. Springer International Publishing, Basel, Schweiz, 582 S.

Hastie, T., R. Tibshirani, und J. Friedman, 2009: *The Elements of Statistical Learning*. Springer-Verlag New York, New York, USA, 745 S.

Hatz, M., 2018: Der Einfluss von mtry auf Random Forests. Masterarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München, München, Deutschland, URL: `https://epub.ub.uni-muenchen.de/59094/1/MA_Hatz.pdf`, abgerufen: 07. Januar 2021.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, **8**, 301–349.

Heimann, D. und M. Kurz, 1985: The Munich hailstorm of July 12, 1984. A discussion of the synoptic situation. *Beiträge zur Physik der Atmosphäre*, **58**, 528–544.

Henze, N., 2010: *Stochastik für Einsteiger: Eine Einführung in die faszinierende Welt des Zufalls*. Vieweg+Teubner Verlag, Wiesbaden, Deutschland, 402 S.

Holton, J. R., 2004: *An Introduction to Dynamic Meteorology*. International Geophysics Series, Elsevier Academic Press, London [u. a.], Vereinigtes Königreich, Burlington, USA, 535 S.

Hosmer, D. W. und S. Lemeshow, 2000: *Applied Logistic Regression*. Wiley–Interscience, New York, USA, 375 S.

Houze, R. A., 1993: *Cloud Dynamics*. Academic Press, San Diego, USA, 570 S.

Houze, R. A., W. Schmid, R. G. Fovell, und H.-H. Schiesser, 1993: Hailstorms in Switzerland: Left movers, right movers, and false hooks. *Monthly Weather Review*, **121**, 3345–3370.

Hunt, B. R., E. J. Kostelich, und I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, **230**, 112–126.

Huntrieser, H., H.-H. Schiesser, W. Schmid, und A. Waldvogel, 1997: Comparison of traditional and newly developed thunderstorm indices for Switzerland. *Weather and Forecasting*, **12**, 108–125.

Höller, H., M. Hagen, P. F. Meischner, V. N. Bringi, und J. Hubbert, 1994: Life cycle and precipitation formation in a hybrid-type hailstorm revealed by polarimetric and Doppler radar measurements. *Journal of the Atmospheric Sciences*, **51**, 2500–2522.

Hübl, J., 2017: Hochwasser Simbach 2016: Dokumentation und Analyse. *Vorsorgender und nachsorgender Hochwasserschutz*, S. Heimerl, Hrsg., Springer Vieweg-Verlag, Wiesbaden, Deutschland, S. 139–150.

Israël, H., 1961: Meteorologie des Gewitters. *Elektrotechnische Zeitschrift*, **8**, 225–231.

James, G., D. Witten, T. Hastie, und R. Tibshirani, 2013: *An Introduction to Statistical Learning*. Springer Verlag New York, New York, USA, 426 S.

James, P. M., B. K. Reichert, und D. Heizenreder, 2018: NowCastMIX: Automatic integrated warnings for severe convection on nowcasting time scales at the German Weather Service. *Weather and Forecasting*, **33**, 1413–1433.

Johns, R. H. und C. A. Doswell, 1992: Severe local storms forecasting. *Weather and Forecasting*, **7**, 588–612.

Josipovic, L., 2020: Interaction of three-dimensional properties in the convective cell development. Masterarbeit, Institut für Atmosphäre und Umwelt, Goethe Universität Frankfurt am Main, Frankfurt am Main, Deutschland.

Kaltenböck, R., G. Diendorfer, und N. Dotzek, 2009: Evaluation of thunderstorm indices from ECMWF analyses, lightning data and severe storm reports. *Atmospheric Research*, **93**, 381–396.

Kalthoff, N., et al., 2009: The impact of convergence zones on the initiation of deep convection: A case study from COPS. *Atmospheric Research*, **93**, 680–694.

Kapsch, M.-L., M. Kunz, R. Vitolo, und T. Economou, 2012: Long-term trends of hail-related weather types in an ensemble of regional climate models using a Bayesian approach. *Journal of Geophysical Research: Atmospheres*, **117**, D15 107.

Kaufman, L. und P. J. Rousseeuw, 1990: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, USA, 342 S.

Kendall, M. und J. D. Gibbons, 1990: *Rank Correlation Methods*. Oxford University Press, New York, USA, 260 S.

Kitzmiller, D. H., W. E. McGovern, und R. F. Saffle, 1995: The WSR-88D severe weather potential algorithm. *Weather and Forecasting*, **10**, 141–159.

Klemp, J. B., 1987: Dynamics of tornadic thunderstorms. *Annual review of fluid mechanics*, **19**, 369–402.

Klemp, J. B. und R. B. Wilhelmson, 1978a: Simulations of right- and left-moving storms produced through storm splitting. *Journal of Atmospheric Sciences*, **35**, 1097–1110.

Klemp, J. B. und R. B. Wilhelmson, 1978b: The simulation of three–dimensional convective storm dynamics. *Journal of the Atmospheric Sciences*, **35**, 1070–1096.

Kober, K. und A. Tafferner, 2009: Tracking and nowcasting of convective cells using remote sensing data from radar and satellite. *Meteorologische Zeitschrift*, **18**, 75–84.

Kraus, H., 2004: *Die Atmosphäre der Erde*. Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, 422 S.

Kuhn, M. und K. Johnson, 2013: *Applied Predictive Modeling*. Springer-Verlag New York, New York, USA, 600 S.

Kunz, M., 2007: The skill of convective parameters and indices to predict isolated and severe thunderstorms. *Natural Hazards and Earth System Sciences*, **7**, 327–342.

Kunz, M., U. Blahak, J. Handwerker, M. Schmidberger, H. J. Punge, S. Mohr, E. Fluck, und K. M. Bedka, 2018: The severe hailstorm in southwest Germany on 28 July 2013: characteristics, impacts and meteorological conditions. *Quarterly Journal of the Royal Meteorological Society*, **144**, 231–250.

Kunz, M., J. Wandel, E. Fluck, S. Baumstark, S. Mohr, und S. Schemm, 2020: Ambient conditions prevailing during hail events in central Europe. *Natural Hazards and Earth System Sciences*, **20**, 1867–1887.

Lang, P., 2001: Cell tracking and warning indicators derived from operational radar products. *Proceedings of the 30th International Conference on Radar Meteorology (München, Deutschland)*, American Meteorological Society, Boston, USA, S. 245–247.

Lang, P., P. Plörer, H. Munier, und J. Riedl, 2003: KONRAD: Konvektionsentwicklung in Radarprodukte – ein operationelles Verfahren zur Analyse von Gewitterzellen und deren Zugbahnen, basierend auf Wetterradarprodukten. Berichte des Deutschen Wetterdienstes 222, Selbstverlag des Deutschen Wetterdienstes, Deutscher Wetterdienst, Offenbach, Deutschland.

Lanicci, J. M. und T. T. Warner, 1991: A synoptic climatology of the elevated mixed-layer inversion over the southern Great Plains in spring. Part I: Structure, dynamics, and seasonal evolution. *Weather and Forecasting*, **6**, 181–197.

Lee, J. W., 2002: Tornado proximity soundings from the NCEP/NCAR reanalysis data. Masterarbeit, School of Meteorology, University of Oklahoma, Norman, USA.

Lee, S. S., 1999: Regularization in skewed binary classification. *Computational Statistics*, **14**, 277–292.

Lemon, L. R. und C. A. Doswell, 1979: Severe thunderstorm evolution and mesocyclone structure as related to tornadogenesis. *Monthly Weather Review*, **107**, 1184–1197.

Li, L., W. Schmid, und J. Joss, 1995: Nowcasting of motion and growth of precipitation with radar over a complex orography. *Journal of Applied Meteorology*, **34**, 1286–1300.

Liaw, A. und M. Wiener, 2018: Package „randomForest“ – Breiman and Cutler’s Random Forests for classification and regression. URL: `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`, abgerufen: 07. Januar 2021.

Lilly, D. K., 1979: The dynamical structure and evolution of thunderstorms and squall lines. *Annual Review of Earth and Planetary Sciences*, **7**, 117–161.

Lin, Y.-L., R. L. Deal, und M. S. Kulie, 1998: Mechanisms of cell regeneration, development, and propagation within a two–dimensional multicell storm. *Journal of the Atmospheric Sciences*, **55**, 1867–1886.

Lin, Y.-L. und L. E. Joyce, 2001: A further study of the mechanisms of cell regeneration, propagation, and development within two–dimensional multicell storms. *Journal of the Atmospheric Sciences*, **58**, 2957–2988.

Ling, C. X. und C. Li, 1998: Data Mining for Direct Marketing: Problems and Solutions. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (New York, USA)*, AAAI Press, New York, USA, S. 73–79.

Lloyd, S. P., 1982: Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 129–137.

Lohmann, U., F. Lüönd, und F. Mahrt, 2016: *An Introduction to Clouds: From the Microscale to Climate*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 391 S.

MacKeen, P. L., H. E. Brooks, und K. L. Elmore, 1999: Radar reflectivity–derived thunderstorm parameters applied to storm longevity forecasting. *Weather and Forecasting*, **14**, 289–295.

MacQueen, J., 1967: Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, USA)*, L. M. Le Cam und J. Neyman, Hrsg., University of California Press, Berkeley, USA, S. 281–297.

Maddox, R. A., 1980: Mesoscale Convective Complexes. *Bulletin of the American Meteorological Society*, **61**, 1374–1387.

Manzato, A. und G. Morgan, 2003: Evaluating the sounding instability with the Lifted Parcel Theory. *Atmospheric Research*, **67–68**, 455–473.

Markowski, P. und Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. John Wiley & Sons, Ltd., Publication, Chichester, Vereinigtes Königreich, 407 S.

Markowski, P. M. und N. Dotzek, 2011: A numerical study of the effects of orography on supercells. *Atmospheric Research*, **100**, 457–478.

Marshall, J. S. und W. M. K. Palmer, 1948: The distribution of raindrops with size. *Journal of Meteorology*, **5**, 165–166.

Marwitz, J. D., 1972a: The structure and motion of severe hailstorms. Part I: Supercell storms. *Journal of Applied Meteorology*, **11**, 166–179.

Marwitz, J. D., 1972b: The structure and motion of severe hailstorms. Part II: Multi-cell storms. *Journal of Applied Meteorology*, **11**, 180–188.

Mason, B., 1971: *The Physics of Clouds*. Oxford University Press, Oxford, Vereinigtes Königreich, 540 S.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291–303.

Mathias, L., V. Ermert, F. Kelemen, P. Ludwig, und J. Pinto, 2017: Synoptic analysis and hindcast of an intense bow echo in western Europe: The 9 June 2014 storm. *Weather and Forecasting*, **32**, 1121–1141.

Mattos, E. V. und L. A. T. Machado, 2011: Cloud-to-ground lightning and Mesoscale Convective Systems. *Atmospheric Research*, **99**, 377–390.

McKelvey, R. D. und W. Zavoina, 1975: A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, **4**, 103–120.

Mecikalski, J., K. M. Bedka, und M. König, 2013: *Best Practice Document, Version 3.0*. EUMETSAT Convection Working Group, Darmstadt, Deutschland.

Mecikalski, J. R., P. D. Watts, und M. Koenig, 2011: Use of Meteosat Second Generation optimal cloud analysis fields for understanding physical attributes of growing cumulus clouds. *Atmospheric Research*, **102**, 175–190.

Mecikalski, J. R., J. K. Williams, C. P. Jewett, D. Ahijevych, A. LeRoy, und J. R. Walker, 2015: Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, **54**, 1039–1059.

Meyer, V. K., H. Höller, und H. D. Betz, 2013: Automated thunderstorm tracking: Utilization of three-dimensional lightning and radar data. *Atmospheric Chemistry and Physics*, **13**, 5137–5150.

Mikuš Jurković, P., N. S. Mahović, und D. Počakal, 2015: Lightning, overshooting top and hail characteristics for strong convective storms in Central Europe. *Atmospheric Research*, **161–162**, 153–168.

Miller, P. W. und T. L. Mote, 2018: Characterizing severe weather potential in synoptically weakly forced thunderstorm environments. *Natural Hazards and Earth System Sciences*, **18**, 1261–1277.

Miller, R. C., 1975: *Notes on Analysis and Severe-Storm Forecasting Procedures of the Air Force Global Weather Central. Vol. 200*. Air Weather Service, Scott Air Force Base, Belleville, USA.

Mohr, S., 2013: *Änderung des Gewitter- und Hagelpotentials im Klimawandel*, Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie, Vol. 58. KIT Scientific Publishing, Karlsruhe, Deutschland, 243 S.

Mohr, S. und M. Kunz, 2013: Recent trends and variabilities of convective parameters relevant for hail events in Germany and Europe. *Atmospheric Research*, **123**, 213–228.

Mohr, S., M. Kunz, A. Richter, und B. Ruck, 2017: Statistical characteristics of convective wind gusts in Germany. *Natural Hazards and Earth System Sciences*, **17**, 957–969.

Mohr, S., J. Wandel, S. Lenggenhager, und O. Martius, 2019: Relationship between atmospheric blocking and warm-season thunderstorms over western and central Europe. *Quarterly Journal of the Royal Meteorological Society*, **145**, 3040–3056.

Mohr, S., et al., 2020: The role of large-scale dynamics in an exceptional sequence of severe thunderstorms in Europe May–June 2018. *Weather and Climate Dynamics*, **1**, 325–348.

Munich Re, 2015: Medieninformationen, Münchener Rückversicherungs-Gesellschaft, München, Deutschland. URL: `https://www.munichre.com/de/unternehmen/media-relations/medieninformationen-und-unternehmensnachrichten/medieninformationen/2015/2015-01-07-naturkatastrophenbilanz-2014-geringere-schaeden-durch-wetterextreme-und-erdbeben.html`, abgerufen: 07. Januar 2021.

Munich Re, 2017: Topics Online, Münchener Rückversicherungs-Gesellschaft, München, Deutschland. URL: `https://www.munichre.com/topics-online/de/climate-change-and-natural-disasters/natural-disasters/floods/rainstorms-europe-2017.html`, abgerufen: 07. Januar 2021.

Munich Re, 2020: Medieninformationen, Münchener Rückversicherungs-Gesellschaft, München, Deutschland. URL: `https://www.munichre.com/de/unternehmen/media-relations/medieninformationen-`

`und-unternehmensnachrichten/medieninformationen/2020/milliardenschaeden-praegen-bilanz-naturkatastrophen-2019.html#1480610047`, abgerufen: 07. Januar 2021.

Murphy, A. H., B. G. Brown, und Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485–501.

Murphy, A. H. und R. L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Nakamura, G. und R. Potthast, 2015: *Inverse Modeling*. IOP Publishing, Bristol, Vereinigtes Königreich, 491 S.

Nestle, R., 1969: Der Tornado vom 10. 7. 1968 im Raum Pforzheim. *Meteorologische Rundschau*, **22**, 1–3.

Nisi, L., P. Ambrosetti, und L. Clementi, 2014: Nowcasting severe convection in the Alpine region: the COALITION approach. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1684–1699.

Nolen, R. H., 1959: A radar pattern associated with tornadoes. *Bulletin of the American Meteorological Society*, **40**, 277–279.

Noppel, H., U. Blahak, A. Seifert, und K. D. Beheng, 2010: Simulations of a hailstorm and the impact of CCN using an advanced two-moment cloud microphysical scheme. *Atmospheric Research*, **96**, 286–301.

Normand, C. W. B., 1931: Graphical indication of humidity in the upper air. *Nature*, **128**, 583.

Norouzi, M., M. Collins, M. A. Johnson, D. J. Fleet, und P. Kohli, 2015: Efficient Non–greedy Optimization of Decision Trees. *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, und R. Garnett, Hrsg., Curran Associates, Inc., Red Hook, USA, S. 1729–1737.

Novák, P. und H. Kyznarová, 2011: Climatology of lightning in the Czech Republic. *Atmospheric Research*, **100**, 318–333.

Orville, H. D. und L. J. Sloan, 1970: A numerical simulation of the life history of a rainstorm. *Journal of the Atmospheric Sciences*, **27**, 1148–1159.

Oshiro, T. M., P. S. Perez, und J. A. Baranauskas, 2012: How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition, 8th International Conference MLDM 2012 (Berlin, Deutschland)*, P. Perner, Hrsg., Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, S. 154–168.

Parker, M. D. und R. H. Johnson, 2000: Organizational modes of midlatitude Mesoscale Convective Systems. *Monthly Weather Review*, **128**, 3413–3436.

Parzen, E., 1962: On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Phillips, D. L., 1962: A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery*, **9**, 84–97.

Piper, D. und M. Kunz, 2017: Spatiotemporal variability of lightning activity in Europe and the relation to the North Atlantic Oscillation teleconnection pattern. *Natural Hazards and Earth System Sciences*, **17**, 1319–1336.

Piper, D., M. Kunz, F. Ehmele, S. Mohr, B. Mühr, A. Kron, und J. Daniell, 2016: Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany – Part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, **16**, 2835–2850.

Pison, G., A. Struyf, und P. J. Rousseeuw, 1999: Displaying a clustering with CLUSPLOT. *Computational Statistics And Data Analysis*, **30**, 381–392.

Prein, A. F. und G. J. Holland, 2018: Global estimates of damaging hail hazard. *Weather and Climate Extremes*, **22**, 10–23.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography*. Developments in Atmospheric Science, Elsevier, Amsterdam, Niederlande, 425 S.

Pruppacher, H. R. und J. D. Klett, 2010: *Microphysics of Clouds and Precipitation*. Kluwer Academic Publishers, Dordrecht, Niederlande, 954 S.

Punge, H. J., K. M. Bedka, M. Kunz, und A. Reinbold, 2017: Hail frequency estimation across Europe based on a combination of overshooting top detections and the ERA-INTERIM reanalysis. *Atmospheric Research*, **198**, 34–43.

Puskeiler, M., M. Kunz, und M. Schmidberger, 2016: Hail statistics for Germany derived from single-polarization radar data. *Atmospheric Research*, **178–179**, 459–470.

Púčik, T., P. Groenemeijer, D. Ryva, und M. Kolar, 2015: Proximity soundings of severe and nonsevere thunderstorms in Central Europe. *Monthly Weather Review*, **143**, 4805–4821.

Rakov, V. A. und M. A. Uman, 2003: *Lightning: Physics and Effects*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 687 S.

Rex, D. F., 1950: Blocking action in the middle troposphere and its effect upon regional climate. *Tellus*, **2**, 275–301.

Rinehardt, R. E. und E. T. Garvey, 1978: Three-dimensional storm motion detection by conventional weather radar. *Nature*, **273**, 287–289.

Ritter, B. und J.-F. Geleyn, 1992: A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Monthly Weather Review*, **120**, 303–325.

Rockel, B., A. Will, und A. Hense, 2008: The regional climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift*, **17**, 347–348.

Romps, D. M., 2017: Exact expression for the lifting condensation level. *Journal of the Atmospheric Sciences*, **74**, 3551–3566.

Rossby, C.-G., 1932: Thermodynamics applied to air mass analysis. Meteorological Papers, Vol. 1, Massachusetts Institute of Technology, Cambridge, USA.

Rotunno, R. und J. Klemp, 1985: On the rotation and propagation of simulated supercell thunderstorms. *Journal of Atmospheric Sciences*, **42**, 271–292.

Rotunno, R., J. B. Klemp, und M. L. Weisman, 1988: A theory for strong, long-lived squall lines. *Journal of Atmospheric Sciences*, **45**, 463–485.

Rousseeuw, P. J., 1987: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Rädler, A. T., P. H. Groenemeijer, E. Faust, R. Sausen, und T. Púčik, 2019: Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Climate and Atmospheric Science*, **2**, 30.

Sauvageot, H., 1992: *Radar Meteorology*. Artech House Publishers, Norwood, USA, 366 S.

Schmid, F., L. Bañon, S. Agersten, A. Atencia, E. de Coning, A. Kann, Y. Wang, und K. Wapler, 2019: Conference Report: Third European Nowcasting Conference. *Meteorologische Zeitschrift*, **28**, 447–450.

Schmidberger, M., 2018: *Hagelgefährdung und Hagelrisiko in Deutschland basierend auf einer Kombination von Radardaten und Versicherungsdaten*, Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie (KIT), Vol. 78. KIT Scientific Publishing, Karlsruhe, Deutschland, 258 S.

Schmidt, K., M. Hagen, H. Höller, E. Richard, und H. Volkert, 2012: Detailed flow, hydrometeor and lightning characteristics of an isolated thunderstorm during COPS. *Atmospheric Chemistry and Physics*, **12**, 6679–6698.

Schraff, C., 1996: *Data Assimilation and Mesoscale Weather Prediction: A Study with a Forecast Model for the Alpine Region*. Swiss Meteorological Institute, Zürich, Schweiz.

Schraff, C., 1997: Mesoscale data assimilation and prediction of low stratus in the Alpine region. *Meteorology and Atmospheric Physics*, **64**, 21–50.

Schraff, C. und R. Hess, 2013: A description of the nonhydrostatic regional COSMO-Model – Part III: Data assimilation. Consortium for Small-Scale Modelling. URL: `http://www.cosmo-model.org/content/model/documentation/core/cosmo_assimilation_5.00.pdf`, abgerufen: 07. Januar 2021.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Periáñez, und R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, **142**, 1453–1472.

Schulz, J.-P. und U. Schättler, 2014: Kurze Beschreibung des Lokal-Modells Europa COSMO-EU (LME) und seiner Datenbanken auf dem Datenserver des DWD. Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/cosmo_eu/cosmo_eu_dbbeschr_201406.pdf?__blob=publicationFile&v=3`, abgerufen: 07. Januar 2021.

Schulz, W., K. Cummins, G. Diendorfer, und M. Dorninger, 2005: Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system. *Journal of Geophysical Research: Atmospheres*, **110**, D09 101.

Schättler, U., G. Doms, und C. Schraff, 2019: A description of the nonhydrostatic regional COSMO-Model – Part VII : User's guide. Consortium for Small-Scale Modelling. URL: `http://www.cosmo-model.org/content/model/documentation/core/cosmo_userguide_5.06a.pdf`, abgerufen: 07. Januar 2021.

Segal, M. R., 2004: *Machine Learning Benchmarks and Random Forest Regression*. Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco, USA.

Seinfeld, J. H. und S. N. Pandis, 2006: *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, New York, USA, 1 152 S.

Selten, F. M., 1995: An efficient description of the dynamics of barotropic flow. *Journal of the Atmospheric Sciences*, **52**, 915–936.

Seltmann, J. E. E. und J. Riedl, 1999: Improved Clutter Treatment Within the German Radar Network: First Results. *COST-75: Advanced Weather Radar Systems, International Seminar, EUR 18567*, C. Collier, Hrsg., Luxembourg, Luxemburg, S. 267–279.

Senf, F. und H. Deneke, 2017: Satellite-based characterization of convective growth and glaciation and its relationship to precipitation formation over Central Europe. *Journal of Applied Meteorology and Climatology*, **56**, 1827–1845.

Senf, F., F. Dietzsch, A. Hünerbein, und H. Deneke, 2015: Characterization of initiation and growth of selected severe convective storms over Central Europe with MSG-SEVIRI. *Journal of Applied Meteorology and Climatology*, **54**, 207–224.

Sherburn, K. D., M. D. Parker, J. R. King, und G. M. Lackmann, 2016: Composite environments of severe and nonsevere high-shear, low-CAPE convective events. *Weather and Forecasting*, **31**, 1899–1927.

Showalter, A. K., 1953: A stability index for thunderstorm forecasting. *Bulletin of the American Meteorological Society*, **34**, 250–252.

Simpson, R. H., 1978: On the computation of equivalent potential temperature. *Monthly Weather Review*, **106**, 124–130.

Spearman, C., 1904: The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72–101.

Stauffer, D. R. und N. L. Seaman, 1990: Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic-scale data. *Monthly Weather Review*, **118**, 1250–1277.

Stephan, K., S. Klink, und C. Schraff, 2008: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Quarterly Journal of the Royal Meteorological Society*, **134**, 1315–1326.

Stephenson, D. B., 2000: Use of the “Odds Ratio” for diagnosing forecast skill. *Weather and Forecasting*, **15**, 221–232.

Strobl, C., A.-L. Boulesteix, A. Zeileis, und T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Student, 1908: The probable error of a mean. *Biometrika*, **6**, 1–25.

Sun, J., et al., 2014: Use of NWP for nowcasting convective precipitation: recent progress and challenges. *Bulletin of the American Meteorological Society*, **95**, 409–426.

SwissRe, 2014: Sigma: Natural catastrophes and man-made disasters in 2013. Swiss Re Economic Research and Consulting, Zürich, Schweiz.

Tang, B. H., V. A. Gensini, und C. R. Homeyer, 2019: Trends in United States large hail environments and observations. *npj Climate and Atmospheric Science*, **2**, 45.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, und H. E. Brooks, 2020: Severe convective storms across Europe and the United States. Part II: ERA5 environments associated with lightning, large hail, severe wind, and tornadoes. *Journal of Climate*, **33**, 10 263–10 286.

Taszarek, M., et al., 2019: A climatology of thunderstorms across Europe from a synthesis of multiple data sources. *Journal of Climate*, **32**, 1813–1837.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, und P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the rapid update cycle. *Weather and Forecasting*, **18**, 1243–1261.

Tiedtke, M., 1989: A comprehensive mass flux scheme for Cumulus parameterization in large-scale models. *Monthly Weather Review*, **117**, 1779–1800.

Tikhonov, A. N., 1963: On the solution of ill-posed problems and the method of regularization. *Doklady Akademii Nauk SSSR*, **151**, 501–504.

Trapp, R. J., 2013: *Mesoscale-Convective Processes in the Atmosphere*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 346 S.

Tsonis, A., 2007: *An Introduction to Atmospheric Thermodynamics*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 187 S.

Ukkonen, P., A. Manzato, und A. Mäkelä, 2017: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *Journal of Applied Meteorology and Climatology*, **56**, 2335–2352.

Vallis, G. K., 2017: *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 946 S.

Van der Laan, M., K. Pollard, und J. Bryan, 2003: A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, **73**, 575–584.

Veall, M. R. und K. Zimmermann, 1996: Pseudo-R2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, **10**, 241–259.

Venables, W. N. und B. D. Ripley, 2013: *Modern Applied Statistics with S-PLUS*. Springer-Verlag Berlin Heidelberg, Berlin [u. a.], Deutschland, 109 S.

Vivekanandan, J., D. S. Zrnic, S. M. Ellis, R. Oye, A. V. Ryzhkov, und J. Straka, 1999: Cloud microphysics retrieval using S-band dual-polarization radar measurements. *Bulletin of the American Meteorological Society*, **80**, 381–388.

Vogel, K., et al., 2017: Die Sturzflut von Braunsbach am 29. Mai 2016 – Entstehung, Ablauf und Schäden eines „Jahrhundertereignisses“. Teil 2: Geomorphologische Prozesse und Schadensanalyse. *Hydrologie und Wasserbewirtschaftung*, **61**, 163–175.

Waldvogel, A., B. Federer, und P. Grimm, 1979: Criteria for the detection of hail cells. *Journal of Applied Meteorology*, **18**, 1521–1525.

Wang, B., W. Shi, und Z. Miao, 2015: Confidence analysis of standard deviational ellipse and its extension into higher dimensional euclidean space. *PloS ONE*, **10**, e0118 537.

Wang, P. K., 2013: *Physics and Dynamics of Clouds and Precipitation*. Cambridge University Press, Cambridge, Vereinigtes Königreich, 452 S.

Wapler, K., 2013: High-resolution climatology of lightning characteristics within Central Europe. *Meteorology and Atmospheric Physics*, **122**, 175–184.

Wapler, K., 2017: The life-cycle of hailstorms: Lightning, radar reflectivity and rotation characteristics. *Atmospheric Research*, **193**, 60–72.

Wapler, K., 2021: Mesocyclonic and non-mesocyclonic convective storms in Germany: Storm characteristics and life-cycle. *Atmospheric Research*, **248**, 105 186.

Wapler, K., L. M. Bañón Peregrín, M. Buzzi, D. Heizenreder, A. Kann, I. Meirold-Mautner, A. Simon, und Y. Wang, 2018: Conference Report 2nd European Nowcasting Conference. *Meteorologische Zeitschrift*, **27**, 81–84.

Wapler, K., F. Harnisch, T. Pardowitz, und F. Senf, 2015: Characterisation and predictability of a strong and a weak forcing severe convective event – A multi-data approach. *Meteorologische Zeitschrift*, **24**, 393–410.

Wapler, K. und P. James, 2015: Thunderstorm occurrence and characteristics in Central Europe under different synoptic conditions. *Atmospheric Research*, **158–159**, 231–244.

Weigl, E., 2015: Radarniederschlag – Prinzip der Niederschlagsbestimmung mit Radar inkl. Umrechnung der Radarreflektivitäten in Momentanwerte des Niederschlages. Deutscher Wetterdienst, Offenbach, Deutschland. URL: `https://www.dwd.de/DE/leistungen/radarniederschlag/rn_info/download_niederschlagsbestimmung.pdf?__blob=publicationFile&v=4`, abgerufen: 07. Januar 2021.

Weisman, M. L., 1992: The role of convectively generated rear-inflow jets in the evolution of long-lived mesoconvective systems. *Journal of Atmospheric Sciences*, **49**, 1826–1847.

Weisman, M. L., 1993: The genesis of severe, long-lived bow echoes. *Journal of Atmospheric Sciences*, **50**, 645–670.

Weisman, M. L. und J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Monthly Weather Review*, **110**, 504–520.

Weisman, M. L., J. B. Klemp, und R. Rotunno, 1988: Structure and evolution of numerically simulated squall lines. *Journal of Atmospheric Sciences*, **45**, 1990–2013.

Weisman, M. L. und R. Rotunno, 2000: The use of vertical wind shear versus helicity in interpreting supercell dynamics. *Journal of the Atmospheric Sciences*, **57**, 1452–1472.

Weisman, M. L. und R. Rotunno, 2004: „A theory for strong long-lived squall lines“ revisited. *Journal of the Atmospheric Sciences*, **61**, 361–382.

Weiss, G. M. und F. Provost, 2001: *The Effect of Class Distribution on Classifier Learning: An Empirical Study*. Department of Computer Science, Rutgers University, New Brunswick, USA.

Werner, M., 2020: Kurzbeschreibung KONRAD3D, Version 1.3.2. Deutscher Wetterdienst, Offenbach, Deutschland (persönliche Kommunikation).

Westermayer, A., P. Groenemeijer, G. Pistotnik, R. Sausen, und E. Faust, 2017: Identification of favorable environments for thunderstorms in reanalysis data. *Meteorologische Zeitschrift*, **26**, 59–70.

Weusthoff, T. und T. Hauf, 2008: Basic characteristics of post-frontal shower precipitation rates. *Meteorologische Zeitschrift*, **17**, 793–805.

Wilhelm, J., 2014: Empirische Orthogonalfunktionen in einem Flachwassermodell auf der Kugel, Bachelorarbeit, Institut für Atmosphäre und Umwelt, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Deutschland.

Wilhelm, J., S. Mohr, H. J. Punge, B. Mühr, M. Schmidberger, J. E. Daniell, K. M. Bedka, und M. Kunz, 2021: Severe thunderstorms with large hail across Germany in June 2019. *Weather*, **76**, 228–237.

Wilhelmson, R., 1974: The life cycle of a thunderstorm in three dimensions. *Journal of the Atmospheric Sciences*, **31**, 1629–1651.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences – Second Edition*. Academic Press, San Diego, USA, 627 S.

Wilks, D. S., 2016: Three new diagnostic verification diagrams. *Meteorological Applications*, **23**, 371–378.

Wilson, J. W., N. A. Crook, C. K. Mueller, J. Sun, und M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bulletin of the American Meteorological Society*, **79**, 2079–2100.

Winterrath, T., C. Brendel, M. Hafer, T. Junghänel, A. Klameth, E. Walawender, E. Weigl, und A. Becker, 2017: Erstellung einer radargestützten Niederschlagsklimatologie, Berichte des Deutschen Wetterdienstes 251, Selbstverlag des Deutschen Wetterdienstes, Deutscher Wetterdienst, Offenbach, Deutschland.

Yeo, I. und R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

Zeidler, E., et al., 2012: *Springer-Taschenbuch der Mathematik: Begründet von I.N. Bronstein und K.A. Semendjaew. Weitergeführt von G. Grosche, V. Ziegler und D. Ziegler. Herausgegeben von E. Zeidler*. Springer Fachmedien Wiesbaden, Wiesbaden, Deutschland, 1 310 S.

Zeng, Y., 2013: *Efficient Radar Forward Operator for Operational Data Assimilation within the COSMO–Model*, Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie, Vol. 60. KIT Scientific Publishing, Karlsruhe, Deutschland, 235 S.

Zeng, Y., U. Blahak, und D. Jerger, 2016: An efficient modular volume-scanning radar forward operator for NWP models: description and coupling to the COSMO model. *Quarterly Journal of the Royal Meteorological Society*, **142**, 3234–3256.

Zinner, T., H. Mannstein, und A. Tafferner, 2008: Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data. *Meteorology and Atmospheric Physics*, **101**, 191–210.

Zöbisch, I., C. Forster, T. Zinner, L. Bugliaro, A. Tafferner, und K. Wapler, 2020: Characteristics of deep moist convection over Germany in multi-source data. *Meteorologische Zeitschrift*, **29**, 393–407.

Zöbisch, I., 2020: Thunderstorms: Life cycle analyses and nowcasting based on multi-source data. Dissertation, Fakultät für Physik, Ludwig-Maximilians-Universität München, München, Deutschland.

A Kurzbeschreibung relevanter konvektiver Indizes

Folgende Kurzbeschreibungen einiger relevanter konvektiver Indizes sind an Kunz (2007), Mohr (2013) sowie an eine Zusammenstellung der *National Oceanic and Atmospheric Administration* (NOAA)[1] angelehnt.

Deep Convective Index (DCI)

Der DCI kombiniert Informationen über die Temperatur und Feuchte im 850 hPa Druckniveau mit der latenten Instabilität, ausgedrückt durch den LI (vgl. Kapitel 2.1.2 und 2.3; Barlow, 1993):

$$\mathrm{DCI} = T_{850\mathrm{hPa}} + \tau_{850\mathrm{hPa}} - \mathrm{LI}\ . \tag{A.1}$$

Werte von $\mathrm{DCI} > 30$ sind Hinweise auf das Potential für das Auftreten schwerer Gewitter.

Vertical Totals (VT)

Der VT ist ein reines Stabilitätsmaß und beschreibt die (bedingte) Instabilität in der mittleren Troposphäre durch die Differenz der Temperaturen im 850 hPa und 500 hPa Druckniveau (Miller, 1975):

$$\mathrm{VT} = T_{850\mathrm{hPa}} - T_{500\mathrm{hPa}}\ . \tag{A.2}$$

[1] `https://www.weather.gov/lmk/indices`

Er ist eng mit der *Lapse Rate* $\mathrm{LR}_{850-500\mathrm{hPa}}$ verknüpft, welche jedoch zusätzlich die unterschiedliche geometrische Schichtdicke bei unterschiedlichen Temperaturen berücksichtigt. Miller und Mote (2018) zeigten beispielsweise, dass der VT besonders in gradientschwachen Wetterlagen im Südosten der USA ein guter Indikator für das Auftreten schwerer Begleiterscheinungen konvektiver Zellen wie Sturmböen oder Hagel sein kann (Unterscheidungstrennwert etwa VT = 25 K).

Total Totals **(TT)**

Als Erweiterung des VT berücksichtigt der TT zusätzlich die Feuchte in der unteren Troposphäre (Miller, 1975):

$$\mathrm{TT} = T_{850\mathrm{hPa}} - 2T_{500\mathrm{hPa}} + \tau_{850\mathrm{hPa}} \ . \qquad \text{(A.3)}$$

Auch der TT ist nach Miller und Mote (2018) ein guter Indikator für das Auftreten schwerer Begleiterscheinungen konvektiver Zellen (Unterscheidungstrennwert etwa TT = 47 K). Huntrieser et al. (1997) zeigten, dass der TT zu den Indizes mit dem besten Unterscheidungsvermögen zwischen Tagen mit und ohne Gewitter in der Schweiz gehört (Unterscheidungstrennwert TT = 45 – 46 K).

K-Index

Der K-Index berücksichtigt im Vergleich zum TT das Maß an Feuchte im 700 hPa Druckniveau und berechnet sich nach George (1960) mit $T^* = T_{850\mathrm{hPa}}$ und $\tau^* = \tau_{850\mathrm{hPa}}$ über

$$\text{K-Index} = T^* - T_{500\mathrm{hPa}} + \tau^* - \left(T_{700\mathrm{hPa}} - \tau_{700\mathrm{hPa}}\right) \ . \qquad \text{(A.4)}$$

Charba (1977) schlug vor, für T^* und τ^* jeweils das arithmetische Mittel der Werte aus dem 850 hPa und dem bodennahen Druckniveau zu verwenden. Trockenlufteinschübe im 700 hPa Druckniveau wie beispielsweise durch die

Advektion einer abgehobenen Mischungsschicht (vgl. Kapitel 2.1.2) können niedrige Werte für den K-Index hervorrufen und ermöglichen dennoch das Auftreten starker konvektiver Zellen, wenn ein genügend starker Hebungsmechanismus vorhanden ist. In den Untersuchungen von Kunz (2007) gehört der K-Index nach Charba (1977) zu den Indizes, die am besten zwischen Tagen mit und ohne Gewitter in Deutschland unterscheiden (Unterscheidungstrennwert K-Index = 35 K).

Vertikaldifferenz von θ_{ps} ($\Delta\theta_{ps}$)

$\Delta\theta_{ps}$ stellt einen Index als Maß für bedingte bzw. potentielle Instabilität dar, der das Potential für starke Fallböen im Abwindbereich einer konvektiven Zelle charakterisiert, und wird daher auch *Wet Mircoburst Index* genannt (Atkins und Wakimoto, 1991). In der Formulierung von Kunz (2007) ist

$$\Delta\theta_{ps} = \theta_{ps,B} - \theta_{ps,300\text{hPa}} \,, \qquad \text{(A.5)}$$

der Index B steht für bodennah. Je höher die Werte von $\Delta\theta_{ps}$ sind, desto höher ist das Fallböenpotential, da die Stärke der Auf- und Abwinde zunimmt, wenn die untere Troposphäre eher warm und feuchtlabil, die obere Troposphäre hingegen eher kalt und trocken ist (vgl. Kapitel 2.1.2).

Showalter Index (SI)

Der SI stellt einen Spezialfall des LI dar (vgl. Kapitel 2.3; Showalter, 1953). Zur Berechnung wird ein Luftpaket angenommen, dessen Ausgangswerte durch das 850 hPa Druckniveau gegeben sind:

$$\text{SI} = T_{500\text{hPa}} - T_{P,500\text{hPa}} \,. \qquad \text{(A.6)}$$

Man beachte, dass in diesem Anhang im Unterschied zu Kapitel 2.1.2 Variablenwerte aus der Umgebung eines Luftpakets ohne den Index U beschrieben werden. Wie für den LI gilt, dass SI < 0 K ein Zeichen für latente Instabilität

für ein solches Luftpaket ist. Im Fall eines Trockenlufteinschubs im Bereich des 850 hPa Druckniveaus kann der SI deutlich höhere Werte anzeigen als solche LI, die auf Luftpaketen aus niedrigeren feuchten Schichten basieren.

KO-Index

Der KO-Index beschreibt die bedingte bzw. potentielle Instabilität (Andersson et al., 1989). Er charakterisiert die mittlere vertikale Änderung der pseudopotentiellen Temperatur in der mittleren Troposphäre über

$$\text{KO-Index} = 0{,}5(\theta_{ps,500\text{hPa}} + \theta_{ps,700\text{hPa}} - 2\theta_{ps,850\text{hPa}}) \,. \qquad \text{(A.7)}$$

Wie für den LI steigt die Instabilität mit sinkenden Werten des KO-Index.

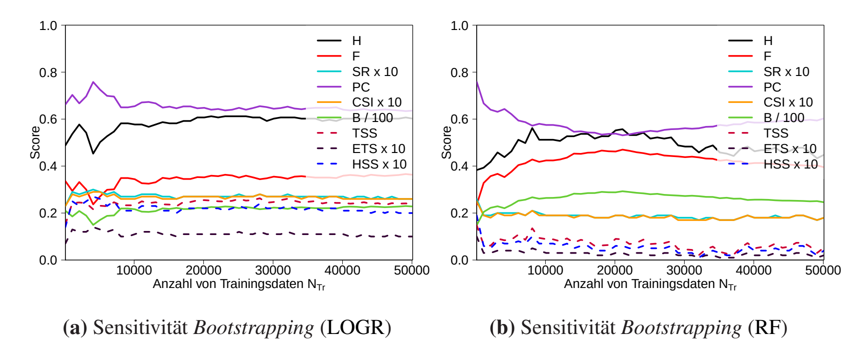
B Sensitivitäten für die Modellstudie U2_0

In den folgenden Abschnitten sind Modellstudien zur Untersuchung der Sensitivität der Vorhersageverfahren bezüglich verschiedener Setup-Parameter am Beispiel der Modellstudie U2_0 erläutert.

Bootstrapping – Größe des Trainingsdatensatzes

Für $\mu_{LOGR} = 0{,}029$ bzw. $\mu_{RF} = 0{,}002$ und einen festen Testdatensatz erfolgt eine Modellstudie zur Sensitivität bezüglich des *Bootstrappings*. Dazu wird jeweils die Größe des Trainingsdatensatzes variiert (Abbildung B.1). Für die logistische Regression ist die Vorhersage nur bis etwa $N_{Tr} = 20\,000$ sensitiv in Bezug auf die Anzahl der Trainingsdaten, während der *Random Forest* die besten Werte für die kategorischen Gütemaße im Bereich zwischen ca. $N_{Tr} = 10\,000$ und $N_{Tr} = 30\,000$ hat. Dies könnte darauf zurückzuführen sein, dass jedes Zellobjekt des Trainingsdatensatzes eine unterschiedliche Gruppe von Entscheidungsbäumen des *Random Forests* durchläuft. Eine Größe des Trainingsdatensatzes $N_{Tr} = \mathcal{O}(f_{Tr}N)$ scheint demnach eine geeignete Wahl zu sein. In den Untersuchungen in diesem Abschnitt sowie in den Kapiteln 6.3 und 6.4 mit einem Ensembleansatz findet daher – sofern nicht anders deklariert – kein *Bootstrapping* statt. Stattdessen dient der potentielle Trainingsdatensatz (alle Zellobjekte, die nicht im Testdatensatz sind) direkt als Trainingsdatensatz. Dieser ist im Fall der Ensemblestudien ohnehin gemäß der Konstruktion des Ensembles für jedes Mitglied verschieden (vgl.

Kapitel 3.6.1 und 6.1.1). Darüber hinaus wird so gewährleistet, dass der maximale Gehalt an Information vorhanden ist: Im Fall des *Bootstrappings* zieht der entsprechende Algorithmus einige Zellobjekte mehrfach, andere aber gar nicht. Letztere sind dort also weder Teil der Trainings- noch der Testdaten.

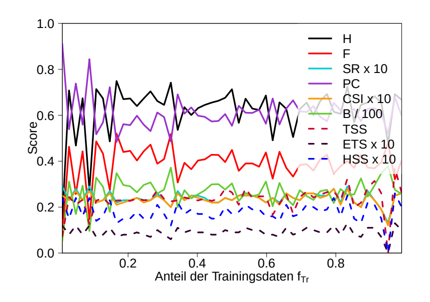


(a) Sensitivität *Bootstrapping* (LOGR) **(b)** Sensitivität *Bootstrapping* (RF)

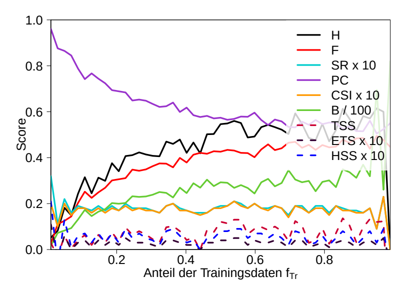
Abbildung B.1: Verschiedene (Skill) Scores basierend auf 51 Modellen mit unterschiedlichen Größen des Trainingsdatensatzes im Intervall $N_{Tr} \in [100; 50\,100]$ in äquidistanten Schritten für (a) die logistische Regression ($\mu_{LOGR} = 0{,}029$) und (b) den *Random Forest* ($\mu_{RF} = 0{,}002$).

Anteil der Trainings- und Testdaten

Für $\mu_{LOGR} = 0{,}029$ bzw. $\mu_{RF} = 0{,}002$ erfolgt eine Modellstudie zur Sensitivität bezüglich des Trainingsdatensatzes. Dazu wird jeweils der Anteil der Trainings- und Testdaten am gesamten Datensatz variiert (Abbildung B.2). Für beide Vorhersagemethoden scheinen Werte für $f_{Tr} \in [0{,}5; 0{,}8]$ am besten geeignet zu sein – dort weisen alle Gütemaße recht stabile Werte auf. Auffällig ist, dass der *Random Forest* zur Stabilisierung der Gütemaße einen deutlich größeren Trainingsdatenanteil benötigt als die logistische Regression. Ein Grund dafür ist möglicherweise die geringe Anzahl von Zellobjekten mit langer Lebensdauer, weswegen der *Random Forest* für niedrige Werte von f_{Tr} fast ausschließlich K-Vorhersagen trifft. Zur Vereinheitlichung erfolgt die Festlegung $f_{Tr} = 0{,}66$ für alle gezeigten Modelle.



(a) Sensitivität Trainingsdatenanteil f_{Tr} (LOGR)



(b) Sensitivität Trainingsdatenanteil f_{Tr} (RF)

Abbildung B.2: Verschiedene (Skill) Scores basierend auf 51 Modellen mit jeweils unterschiedlichem Anteil der Trainings- und Testdaten am gesamten Datensatz im Intervall $f_{Tr} \in [0{,}01\,;0{,}99]$ in äquidistanten Schritten für (a) die logistische Regression ($\mu_{LOGR} = 0{,}029$) und (b) den *Random Forest* ($\mu_{RF} = 0{,}002$).

Wahl des Klassentrennwerts

Beide Vorhersagemethoden werden im Anschluss in einer Modellstudie zur Sensitivität bezüglich des Klassentrennwerts der Lebensdauer mit verschiedenen Werten für τ für jeweils ein exemplarisches Modell mit festem Trainings- und Testdatensatz ausgetestet (Tabelle B.1; Abbildung B.3). Die halbe Breite des Übergangsbereichs ist wie in der Modellstudie U2_0 durch $\tau' = 15\,\text{min}$ gegeben. Die Intervalle für den Entscheidungstrennwert μ_{LOGR} sind für jeden Klassentrennwert unterschiedlich, um eine möglichst glatte ROC-Kurve zu erhalten. Im Fall des *Random Forests* ist wie in U2_0 $\mu_{RF} \in [0{,}001\,;0{,}101]$ eine geeignete Wahl, da der *Random Forest* mit 1 000 Bäumen Werte von $\mu_{RF} < 0{,}001$ nicht abbilden kann (vgl. Kapitel 3.5.2).

Die Vorhersagen der logistischen Regression erreichen umso höhere Werte für die *AUC*, je größer τ ist. Allerdings zeigt sich für $\tau > 90\,\text{min}$ keine weitere Verbesserung. Auch die Realisierungen des *Random Forests* deuten an, dass die Vorhersagen für größere τ bessere Scores erwarten lassen. Der

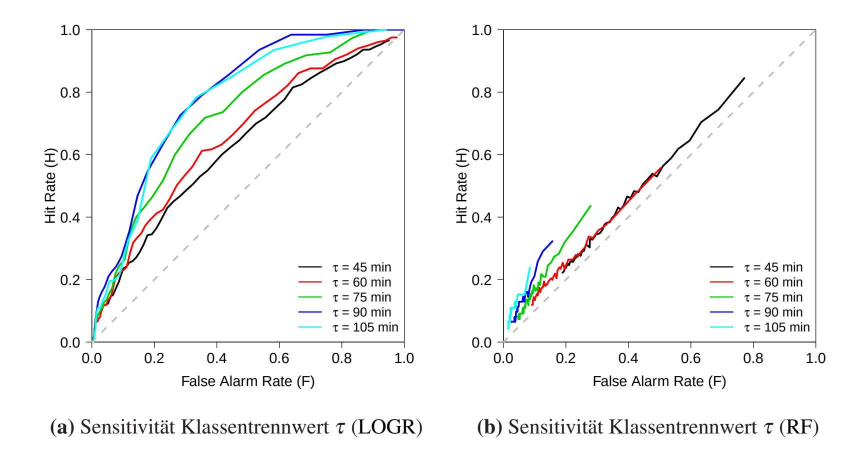
Tabelle B.1: Anzahl von Zellobjekten im gesamten Datensatz mit langer Lebensdauer $N_L > \tau$, entsprechende Klassenverhältnisse ρ_K sowie AUC für ein beispielhaftes Modell der logistischen Regression mit verschiedenen Klassentrennwerten τ.

Klassentrennwert $\tau \rightarrow$ Maß $\downarrow$	45 min	60 min	75 min	90 min	105 min
N_L	2 280	1 096	560	309	189
ρ_K	0,063	0,029	0,014	0,008	0,005
AUC (LOGR)	0,632	0,659	0,724	0,787	0,775

Unterschied zwischen den ROC-Kurven ist generell höher als die Variabilität, die sich durch die Wahl unterschiedlicher Trainings- und Testdaten ergibt (vgl. Abbildung 6.2) und damit signifikant. Aufgrund des Übergangsbereichs mit einer halben Breite von $\tau' = 15$ min gehen allgemein für $\tau = 60$ min rund 200, für $\tau = 75$ min noch rund 100 und $\tau = 90$ min noch rund 60 Zellobjekte in die Evaluation der einzelnen Realisierungen ein.

Wahl des Übergangsbereichs für den Klassentrennwert

Beide Vorhersagemethoden werden nun in einer Modellstudie zur Sensitivität bezüglich des Übergangsbereichs für den Klassentrennwert der Lebensdauer für fünf verschiedene Werte von τ' mit dem Klassentrennwert $\tau = 60$ min ausgetestet (Abbildung B.4). Dazu wird jeweils ein Ensemble aus 51 Modellen mit festen Entscheidungstrennwerten $\mu_{LOGR} = 0{,}029$ bzw. $\mu_{RF} = 0{,}002$ aufgesetzt. Je größer der Übergangsbereich ist, desto höher sind die Werte für die *ACC* und desto niedriger jene für den *BS* für Zellobjekte mit langer Lebensdauer. Dabei ändert sich die *ACC* nur wenig, während der *BS* eine deutliche Abnahme anzeigt. Dies lässt sich dadurch erklären, dass durch die Vergrößerung des symmetrischen Übergangsbereichs ein nicht unerheblicher



(a) Sensitivität Klassentrennwert τ (LOGR) **(b)** Sensitivität Klassentrennwert τ (RF)

Abbildung B.3: ROC-Kurven basierend auf 51 Realisierungen mit unterschiedlichen Klassentrennwerten für ein beispielhaftes Modell – für $\tau = 60\,\text{min}$ wie eines der Modelle in Abbildung 6.2a+b – (a) der logistischen Regression und (b) des *Random Forests*.

Anteil von Zellobjekten mit langer Lebensdauer nicht in die Evaluierung eingeht. Gerade Zellobjekte mit einer langen Lebensdauer nahe am Klassentrennwert τ sind deutlich stärker vertreten als solche mit sehr langer Lebensdauer (vgl. Kapitel 6.1.2).

Hier wird deutlich, dass die Wahl des Übergangsbereichs einen entscheidenden Einfluss auf die Gütemaße hat. Die Wahl $\tau = 60\,\text{min}$ mit $\tau' = 15\,\text{min}$ für U2_0 sowie die Modellstudien in Kapitel 6.3.1 stellt aus statistischer Sicht einen guten Kompromiss dar, der sowohl eine klare Separation der beiden Klassen als auch eine adäquate Anzahl von Zellobjekten mit langer Lebensdauer im Testdatensatz, die in die Evaluation einer einzelnen Realisierung oder eines einzelnen Ensemblemitglieds eingehen (nämlich rund 200, s. o.), erreicht. Aus meteorologischer Sicht erscheint diese Kombination ebenfalls als sinnvoll, spiegelt doch die Evaluation von Zellobjekten mit einer Lebensdauer von weniger als 45 min eher die Auswertung von nur wenig organisierten Einzelzellen, und die Evaluation von Objekten mit einer

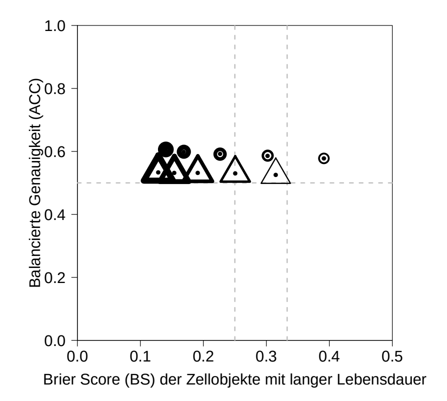


Abbildung B.4: Analog zu Abbildung 6.6, nur für die Untersuchung der Sensitivität bezüglich der halben Breite des Übergangsbereichs τ' des Klassentrennwerts τ. Je dicker die Symbole sind, desto größer ist der Übergangsbereich für $\tau = 60$ min ($\tau' = 0, 5, 10, 15, 20$ min).

Lebensdauer von mehr als 75 min eher diejenige von organisierter Konvektion wider (vgl. Kapitel 2.2). Die Auswertungen der in Kapitel 6.3.1 gezeigten Modellstudien müssen daher mit dem Wissen um die jeweiligen Sensitivitäten interpretiert werden.

Wahl der Größe des Ensembles

Auch die Wahl der Ensemblegröße zeigt einen, obgleich auch geringen Einfluss auf die Evaluation. Weniger als 51 Ensemblemitglieder sind für $f_{Tr} = 0{,}66$ ohnehin nicht zweckmäßig, da sonst viele Zellobjekte nur in wenigen Testdatensätzen vorkommen und eine probabilistische Auswertung nicht sinnvoll ist. Vergrößert man das Ensemble aus U2_0 um 20 oder 40 Mitglieder, so unterscheiden sich beispielsweise der *BS* für die Zellobjekte mit langer Lebensdauer, die *ACC* sowie die mittlere Schwankungsbreite des Ensembles $\hat{\overline{\sigma}}_{ens}$ jeweils um weniger als 2 % von den Werten, die man mit 51 Ensemblemitgliedern erhält (nicht gezeigt). Ebenso gibt es kaum einen Unterschied in

den Verteilungen der Vorhersagen aus Abbildung 6.5. Dieses Ergebnis lässt sich auch für die Modellstudien aus den Kapiteln 6.3 und 6.4 feststellen, sodass in allen Modellstudien 51 Ensemblemitglieder verwendet werden.

Wahl des Verfahrens für die Transformation der Werte der Prädiktoren

Das Verfahren zur Transformation der Werte der Prädiktoren (vgl. Kapitel 3.5.1 und 6.1.1) stellt sich in U2_0 als nicht entscheidend heraus. Sowohl DLS als auch $\mathrm{LI}_{100\mathrm{hPa}}$ weichen beide nicht allzu stark von der Normalverteilung ab. So unterscheiden sich die ROC-Kurven, die man in U2_0 nach Durchführung der Kombination einer z- und einer Yeo-Johnson-Transformation erhält, kaum von denen, die man nach einer alleinigen z-Transformation bzw. gänzlich ohne Transformation erhält (für die logistische Regression beispielsweise ändert sich die *AUC* um weniger als 0,5 %; nicht gezeigt). Da man die Effekt-Koeffizienten der logistischen Regression bzw. die Koeffizienten der linearen Regression (vgl. Kapitel 3.3.1) direkt in Bezug zur Wichtigkeit der Prädiktoren setzen kann, wenn eine z-Transformation der Werte der Prädiktoren erfolgt, soll dies für die weiteren Untersuchungen beibehalten werden. Ebenso ist die Anwendung der Yeo-Johnson-Transformation insbesondere für Prädiktoren sinnvoll, deren Verteilung stark von der Normalverteilung abweicht. Sie wird daher in allen Modellstudien auf alle Prädiktoren angewendet.

Resampling zur Balancierung des Trainingsdatensatzes

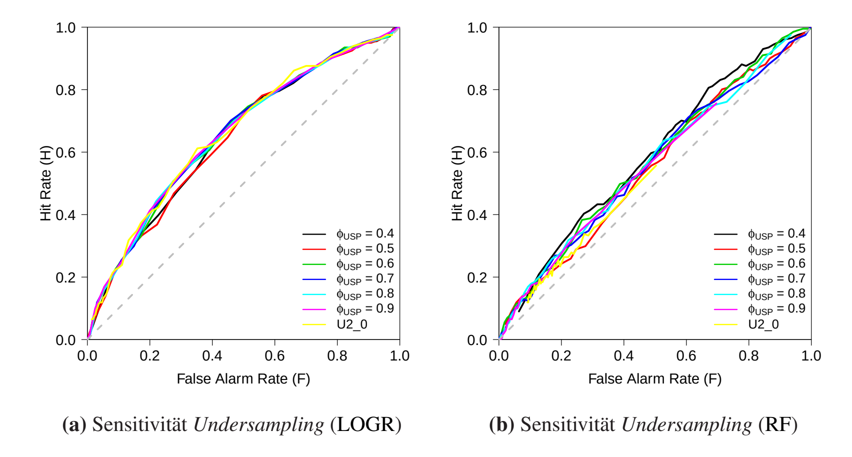
Das *Undersampling* und das *Oversampling* können zur Balancierung des Trainingsdatensatzes beitragen, d. h. das Klassenverhältnis ρ_K im Vergleich zu U2_0 vergrößern. Dabei erhofft man sich eine Verbesserung der Güte der Vorhersagen für einen balancierten Datensatz (vgl. Kapitel 3.5.2 und 6.1.1).

Tabelle B.2: Anzahl von Zellobjekten N'_{Tr} sowie von Zellobjekten mit langer Lebensdauer $N'_{L,Tr}$ im modifizierten Trainingsdatensatz, dortige Klassenverhältnisse $\rho'_{K,Tr}$ sowie AUC für ein beispielhaftes Modell der logistischen Regression sowie des *Random Forests* mit verschiedenen Werten für den Balanceparameter ϕ_{USP} bei der Anwendung eines *Undersamplings*.

$\phi_{USP} \rightarrow$ Maß $\downarrow$	0,4	0,5	0,6	0,7	0,8	0,9	U2_0
N'_{Tr}	136	322	514	1 298	2 623	6 702	25 000
$N'_{L,Tr}$	98	202	297	566	675	675	675
$\rho'_{K,Tr}$	2,58	1,68	1,37	0,77	0,35	0,10	0,03
AUC (LOGR)	0,65	0,64	0,66	0,66	0,66	0,66	0,66
AUC (RF)	0,59	0,55	0,57	0,56	0,57	0,56	—

Undersampling

Zunächst wird das *Undersampling* untersucht (Tabelle B.2, Abbildung B.5). Die Entscheidungstrennwerte μ_{LOGR} und μ_{RF} variieren im Intervall [0,01 ; 0,99]. Durch die Wahl des Balanceparameters ϕ_{USP} hat man direkt Einfluss auf die Gesamtanzahl von Zellobjekten im modifizierten Trainingsdatensatz N'_{Tr}, die dortige Anzahl von Zellobjekten mit langer Lebensdauer $N'_{L,Tr}$ sowie das dortige Klassenverhältnis $\rho'_{K,Tr}$. Offensichtlich führen kleine Werte von ϕ_{USP} zu sehr kleinen Trainingsdatensätzen, in denen mehr Zellobjekte mit langer als mit kurzer Lebensdauer zu finden sind. Eine adäquate Wahl für den Balanceparameter liegt in etwa im Intervall $\phi_{USP} \in [0{,}6 ; 0{,}7]$: Hier ist der Trainingsdatensatz hinreichend groß und die Anteile von Zellobjekten mit kurzer und langer Lebensdauer sind recht gut ausbalanciert. Die AUC-Werte weisen darauf hin, dass die Vorhersagegüte für die logistische Regression derjenigen aus U2_0 nicht nachsteht, jedoch auch nicht zugenommen hat.



(a) Sensitivität *Undersampling* (LOGR) **(b)** Sensitivität *Undersampling* (RF)

Abbildung B.5: ROC-Kurven basierend auf 51 Realisierungen mit unterschiedlichen Entscheidungstrennwerten für ein beispielhaftes Modell (a) der logistischen Regression und (b) des *Random Forests* mit *Undersampling* des Trainingsdatensatzes (für $\phi_{USP} = 1{,}0$ wie eines der Modelle in Abbildung 6.2a+b).

Dieses Resultat steht nicht im Widerspruch zum Ergebnis aus der Untersuchung der Größe des Trainingsdatensatzes. Während dort eine Verkleinerung des Datensatzes mit einer Reduzierung von Zellobjekten gemäß der allgemeinen Verteilung der Werte des Prädiktanden einhergeht, entfernt das *Undersampling* hauptsächlich Zellobjekte mit kurzer Lebensdauer aus dem Trainingsdatensatz (für $\phi_{USP} = 0{,}7$ beispielsweise knapp 97 % aller Objekte mit kurzer Lebensdauer aus U2_0), während die Anzahl von Zellobjekten mit langer Lebensdauer nur wenig sinkt (für $\phi_{USP} = 0{,}7$ nur um etwa 16 %). Offensichtlich geht trotz der starken Verkleinerung des originalen Trainingsdatensatzes nur wenig von dessen Informationsgehalt durch das *Undersampling* verloren.

Oversampling

In ähnlicher Weise geschieht die Untersuchung des *Oversamplings*. Hier führen hohe Werte von ϕ_{OSP} zu größeren Trainingsdatensätzen

– beispielsweise finden sich dort für $\phi_{OSP} = 0{,}9$ genau $N'_{Tr} = 35\,115$ Zellobjekte mit einem Klassenverhältnis $\rho'_{K,Tr} = 35{,}4\,\%$ und $N'_{L,Tr} = 9\,187$ Zellobjekte mit langer Lebensdauer – die meisten davon natürlich fiktiv. Die Wahl der *Oversampling*-Methode ist für die Qualität der Vorhersage nicht maßgebend. Sowohl mit SMOTE als auch mit dem Gauss'schen Rauschen und dem zufälligen *Oversampling* mit beliebigem ϕ_{OSP} ähneln die ROC-Kurven derjenigen aus U2_0 stark (nicht gezeigt). Auch hier gibt es keine Verbesserung der Gütemaße. Es sei angemerkt, dass nur für ϕ_{OSP} nahe 1 und einer dementsprechend hohen Anzahl von fiktiven Zellobjekten $\rho'_{K,Tr}$ Werte nahe 1 erreicht. Insgesamt ist das *Oversampling* alleine daher nicht von Nutzen.

Kombination aus *Undersampling* und *Oversampling*

Eine Kombination aus *Undersampling* und *Oversampling* ist am sinnigsten, wenn zunächst das *Undersampling* und anschließend das *Oversampling* erfolgt (vgl. Kapitel 3.5.2). Als erstes wird demnach die Anzahl von Zellobjekten mit kurzer Lebensdauer verringert und im Anschluss die Anzahl von Zellobjekten mit langer Lebensdauer durch die Einführung von fiktiven Zellobjekten erhöht. Mit der hier angewendeten Vorgehensweise erhält man etwa für beliebige $\phi_{USP} > 0{,}7$ ähnliche Klassenverhältnisse $\rho'_{K,Tr}$ – je nach Wahl der minimalen Anzahl von Zellobjekten einer beobachteten Lebensdauer $N_{I,min}$, die erreicht werden muss, sodass das *Oversampling* diese berücksichtigt (vgl. Kapitel 3.5.2).

Beispielhaft wird $N_{I,min} = 15$ gewählt und ϕ_{USP} im gleichen Intervall variiert wie bei den Untersuchungen zum *Undersampling* (Tabelle B.3). Damit liegt das Klassenverhältnis im Trainingsdatensatz beispielsweise für $\phi_{USP} = 0{,}7$ bei ca. $\rho'_{K,Tr} \approx 127\,\%$. Für kleine Werte von ϕ_{USP} geht das Verfahren in ein reines *Undersampling* über, für große ϕ_{USP} in ein reines *Oversampling*. Bei den Werten für die *AUC* des *Random Forests* fällt auf, dass die Werte geringfügig höher als beim reinen *Undersampling* liegen (vgl. Tabelle B.2). Für den *Random Forest* liegt der maximale Unterschied bei 2,2, für die logistische

Tabelle B.3: Anzahl von Zellobjekten N'_{Tr} sowie von Zellobjekten mit langer Lebensdauer $N'_{L,Tr}$ im Trainingsdatensatz, dortige Klassenverhältnisse $\rho'_{K,Tr}$ sowie *AUC* für ein beispielhaftes Modell der logistischen Regression sowie des *Random Forests* mit verschiedenen Werten für den Balanceparameter ϕ_{USP} bei einer Kombination von *Undersampling* und *Oversampling* (Gauss'sches Rauschen) mit $N_{I,min} = 15$.

$\phi_{USP} \rightarrow$ Maß $\downarrow$	0,4	0,5	0,6	0,7	0,8	0,9	U2_0
N'_{Tr}	136	323	523	1 662	4 503	16 810	25 000
$N'_{L,Tr}$	98	202	306	930	2 478	9 175	675
$\rho'_{K,Tr}$	2,58	1,67	1,41	1,27	1,22	1,20	0,03
AUC (LOGR)	0,65	0,64	0,66	0,66	0,66	0,66	0,66
AUC (RF)	0,59	0,55	0,57	0,57	0,59	0,57	—

Regression bei 0,3 Prozentpunkten. Man erhält diese höheren Werte, indem man als *Oversampling*-Methode das Gauss'sche Rauschen einsetzt. Auch die *AUC*-Werte unter Verwendung von SMOTE zeigen eine ähnliche Verbesserung, wohingegen zufälliges *Oversampling* keine Verbesserung liefert (nicht gezeigt).

Für Klassifikationsverfahren empfiehlt sich daher die Anwendung eines *Resamplings* bzw. eine Kombination aus *Undersampling* und *Oversampling* nur bedingt, da die Wahl des Entscheidungstrennwerts μ ebenso die Balancierung der Vorhersagen steuern kann. Für den *Random Forest* ist ein *Resampling* dann nützlich, wenn es ermöglicht, die notwendige Anzahl an Bäumen N_{Baum} deutlich zu reduzieren und/oder der *Random Forest* gleich gute oder gar bessere Vorhersagen trifft (vgl. Kapitel 3.5.2).

Spezifische Tuningparameter für den *Random Forest*

Der offensichtlichste frei wählbare Parameter bei der Generierung eines *Random Forests* ist die Anzahl von Entscheidungsbäumen, die dem *Random Forest* zugrunde liegen (vgl. Kapitel 3.4.3). Ähnlich der relativen Häufigkeit des Ergebnisses eines Zufallsexperiments stabilisiert sich die Vorhersage eines *Random Forests* mit steigender Anzahl von Entscheidungsbäumen N_{Baum} (vgl. Gesetz der großen Zahlen; z. B. Henze, 2010). Anschaulich gesprochen nähern sich die ROC-Kurven für $N_{Baum} \rightarrow \infty$ einer asymptotischen Kurve an.

Oshiro et al. (2012) postulierten auf der Basis eines experimentellen Setups mit 29 verschiedenen Datensätzen, dass eine Zahl zwischen 64 und 128 Entscheidungsbäumen bezüglich verschiedener Gütemaße allgemein empfehlenswert sei. Gerade bei einer sehr großen Anzahl von Prädiktoren könne es jedoch nützlich sein eine größere Zahl zu verwenden. Des Weiteren sind, wie in U2_0 gesehen, große Werte für N_{Baum} unabdingbar, wenn der Datensatz sehr unbalanciert ist.

Bei der Variation von N_{Baum} wird deutlich, dass für ein Setup mit zwei Prädiktoren ca. 75 Entscheidungsbäume bereits ausreichend sind, wenn eine Kombination aus *Undersampling* und *Oversampling* für den Trainingsdatensatz Anwendung findet (Abbildung B.6). Der Unterschied zwischen der ROC-Kurve mit 75 und derjenigen mit 2 000 Entscheidungsbäumen liegt hier lediglich bei 1,3 %. 2 000 Entscheidungsbäume gelten in diesem Fall nach genauer Betrachtung in sehr guter Näherung als Referenz ($N_{Baum} \rightarrow \infty$). Der Unterschied zwischen den ROC-Kurven ist definiert als das Mittel der Abstände der Punkte der 51 Realisierungen zu den Punkten mit $N_{Baum} = 2\,000$, welche die ROC-Kurven im F-H-Raum aufspannen.

Ein weiterer frei wählbarer Parameter bei der Generierung eines *Random Forests* ist die Anzahl von Prädiktoren, die angibt, wie viele Prädiktoren bei jedem Split innerhalb eines Entscheidungsbaums auf den bestmöglichen Split getestet werden sollen (vgl. Kapitel 3.4; z. B. Hatz, 2018). Die Unterschiede

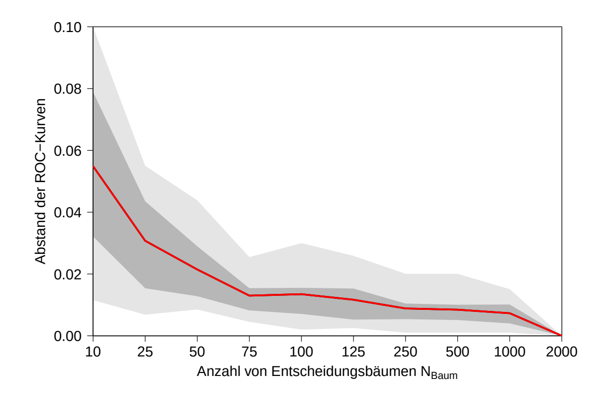


Abbildung B.6: Abstand der ROC-Kurven für unterschiedliche Anzahlen von Entscheidungsbäumen N_{Baum} in jeweils 51 *Random Forest*-Realisierungen mit unterschiedlichem $\mu_{RF} \in [0{,}01\,;0{,}99]$ bei einer Kombination aus *Undersampling* und *Oversampling* (ϕ_{USP} = 0,7) zu der als Referenz angesehenen ROC-Kurve mit 2 000 Entscheidungsbäumen. Die rote Kurve gibt das arithmetische Mittel über die 51 Modelle an; der dunkelgraue Bereich stellt den Interquartilsbereich dar; der hellgraue Bereich markiert das 5. und 95. Perzentil.

zwischen ähnlichen Werten für N_{split} sind – bezogen auf die hier vorliegende Arbeit – in der Regel nicht von Bedeutung und bedürfen keiner tieferen Diskussion. Soweit nicht anders angegeben werden im Folgenden die in Kapitel 3.4.3 genannten Standardwerte verwendet. Ist $1 < N_{po} < 6$, so ist $N_{split} = 2$ eine geeignete Wahl.

Über die beiden genannten Parameter hinaus gibt es einige weitere Parameter, welche die Charakteristik eines *Random Forests* beschreiben, wie z. B. solche, die das Wachstum der Entscheidungsbäume beeinflussen (vgl. Kapitel 3.4). Diese Parameter bleiben in dieser Arbeit unangetastet und es werden die von Liaw und Wiener (2018) vorgegebenen Standardwerte verwendet. Nähere Informationen finden sich beispielsweise dort wie auch bei Breiman (2001).

C Abschätzung der Variabilität der Lebensdauer im Parabelmodell

Für den in Kapitel 5.1.2 vorgestellten Parabelansatz erhält man durch Umformulieren von Gleichung (5.2):

$$T_Z = \frac{4c_A t^2}{4c_A t - A_Z(t) + \mu_A} \equiv \frac{\zeta(t)}{\eta(t, A_Z(t))} \, . \tag{C.1}$$

Gleichung (C.1) liefert nur für $\mu_A \leq A_Z(t) < A_{Z,krit}(t) = 4c_A t + \mu_A$ sinnvolle Werte (vgl. Kapitel 5.1.2). Für $A_Z(t) < \mu_A$ ist $T_Z < t$ und bei $A_Z(t) = A_{Z,krit}(t)$ findet sich eine Polstelle mit $T_Z \rightarrow \pm\infty$ für $A_Z(t) \rightarrow A_{Z,krit}(t)$. Die maximale zu erwartende Zellfläche kann mittels (5.2) auf

$$A_{Z,max}^{(T_Z)} = \mu_A + c_A T_Z \tag{C.2}$$

abgeschätzt werden.

Durch die zeitliche Auflösung der Beobachtungen von 5 min ergeben sich aufgrund der Bestimmungsmethodik der Zellfläche verschiedene Unsicherheiten: Möglicherweise wird ein Teil der detektierten konvektiven Zelle von sich selbst oder von anderen Zellen abgeschattet, oder der Reflektivitätsfaktor liegt an einigen Gitterpunkten im Radarbild knapp unter 46 dBZ, sodass diese nicht zum Zellobjekt dazuzählen. Zur Abschätzung dieser Unsicherheiten für die zu erwartende Lebensdauer bietet sich entweder die Variation des beobachteten Zeitpunkts um einige Minuten oder die Variation der beobachteten Zellfläche zu einem Zeitpunkt an. Für letztere gilt beispielsweise folgende

Tabelle C.1: Beispielwerte der zu erwartenden Lebensdauer T_Z sowie Unsicherheiten derer ΔT_Z für verschiedene Werte der Variation der Zellfläche ΔA_Z, berechnet mit der exakten Variante von Gleichung (C.3). In den Spalten 2 bis 4 sind die Werte für $A_Z = 30\,\text{km}^2$ zu verschiedenen Zellaltern t dargestellt, in den Spalten 5 und 6 solche für $A_Z = 45\,\text{km}^2$.

A_Z (km^2) $\rightarrow$	30			45	
t (min) $\rightarrow$	15	30	60	30	60
T_Z (min) $\rightarrow$	29,4	39,7	68,4	90,4	90,1
ΔA_Z (km^2) $\downarrow$	ΔT$_Z$ (min) $\downarrow$				
−10	−15,4	−10,8	−9,5	−41,5	−15,8
−6	−11,7	−7,3	−6,0	−30,5	−10,2
−2	−5,3	−2,8	−2,1	−13,1	−3,7
+2	8,3	3,2	2,3	18,5	4,0
+6	57,9	11,5	7,3	94,1	13,1
+10	—	23,7	13,1	513,4	24,2

Abschätzung:

$$\begin{aligned} T_Z + \Delta T_Z &= \frac{\zeta}{\eta - \Delta A_Z} \\ \Longleftrightarrow \Delta T_Z &= \zeta \left(\frac{1}{\eta - \Delta A_Z} - \frac{1}{\eta} \right) \\ &\approx \frac{\zeta}{\eta} \left[\frac{\Delta A_Z}{\eta} + \left(\frac{\Delta A_Z}{\eta} \right)^2 + \mathcal{O} \left(\frac{\Delta A_Z}{\eta} \right)^3 \right] . \end{aligned} \tag{C.3}$$

Die Unsicherheit bezüglich der Amplitude der Entwicklung der Zellfläche kann dann einfach über $\Delta \mathscr{A} = c_A \Delta T_Z$ bestimmt werden. Die Unsicherheit bezüglich der zu erwartenden Lebensdauer ist größer, je jünger das Zellobjekt ist (Tabelle C.1). Hier liegen die Kurven der Parabelschar in Abbildung 5.8a sehr dicht beisammen. Für ein Zellobjekt, das beispielsweise nach 17 min

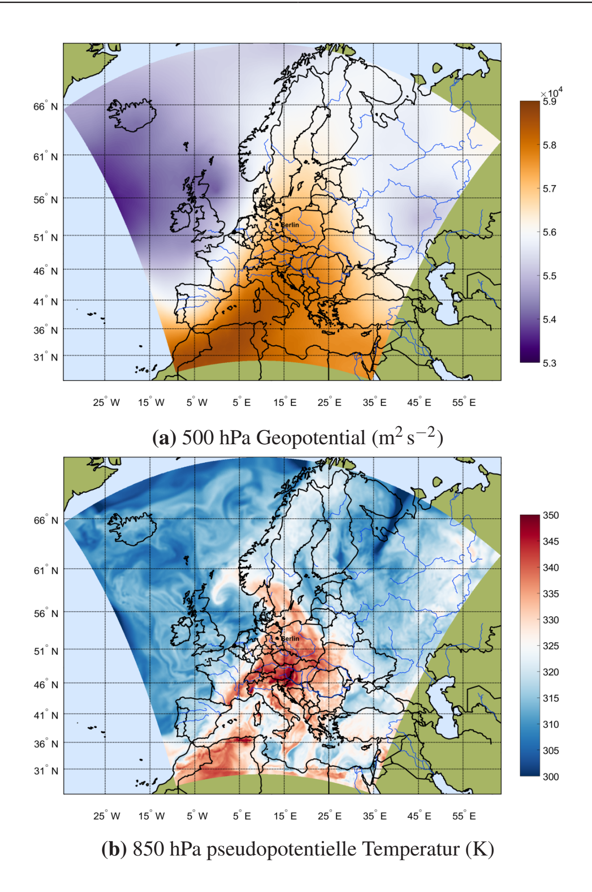
eine Größe von 30 km^2 erreicht hat, prognostiziert das Parabelmodell deterministisch eine verbleibende Lebenszeit von 14,4 min ($T_Z = 29{,}4$ min). Ist ein Zellobjekt zum gleichen Zeitpunkt 36 km^2 groß ($\Delta A_Z = 6\,\mathrm{km}^2$), so lautet die Prognose 57,9 min länger, also 72,3 min verbleibende Lebenszeit. Betrachtet man zwei Zellobjekte nach 30 min, ist die Entwicklung des einen Objekts mit einer Fläche von 45 km^2 deutlich unsicherer als die des anderen mit einer Fläche von 30 km^2. Die Fläche des letzteren wird sich in den folgenden Minuten sehr wahrscheinlich rasch verkleinern.

Die Gleichungen (C.1) und (C.3) gelten darüber hinaus für den erweiterten Parabelansatz aus Kapitel 5.3.1 mit der Ersetzung $c_A = c_A(u)$, der die Abhängigkeit der Parabelschar von einer Umgebungsvariablen berücksichtigt. Für $t = 30$ min und $A_Z = 45\,\mathrm{km}^2$ ergibt sich ohne Berücksichtigung einer Umgebungsvariablen gemäß Tabelle C.1 eine zu erwartende Lebensdauer von $T_Z \approx 90$ min. In Abhängigkeit vom $\mathrm{LI}_{100\mathrm{hPa}}$ findet man hingegen folgende Werte für T_Z:

$$\begin{aligned}
\mathrm{LI}_{100\mathrm{hPa}} = 3\,\mathrm{K} &\implies T_Z \approx 118{,}7\,\mathrm{min} \\
\mathrm{LI}_{100\mathrm{hPa}} = 1\,\mathrm{K} &\implies T_Z \approx 81{,}6\,\mathrm{min} \\
\mathrm{LI}_{100\mathrm{hPa}} = -1\,\mathrm{K} &\implies T_Z \approx 66{,}4\,\mathrm{min} \\
\mathrm{LI}_{100\mathrm{hPa}} = -3\,\mathrm{K} &\implies T_Z \approx 58{,}1\,\mathrm{min} \\
\mathrm{LI}_{100\mathrm{hPa}} = -5\,\mathrm{K} &\implies T_Z \approx 52{,}9\,\mathrm{min} \qquad \text{(C.4)}
\end{aligned}$$

Je niedriger der $\mathrm{LI}_{100\mathrm{hPa}}$, desto höher ist die Erwartung für die Entwicklung einer Zellfläche von mehr als 45 km^2 nach der ersten halben Stunde. Daher sinkt die zu erwartende Lebensdauer, wenn das Objekt trotz guter Umgebungsbedingungen (hier: LI) in der ersten halben Stunde nicht allzu stark gewachsen ist.

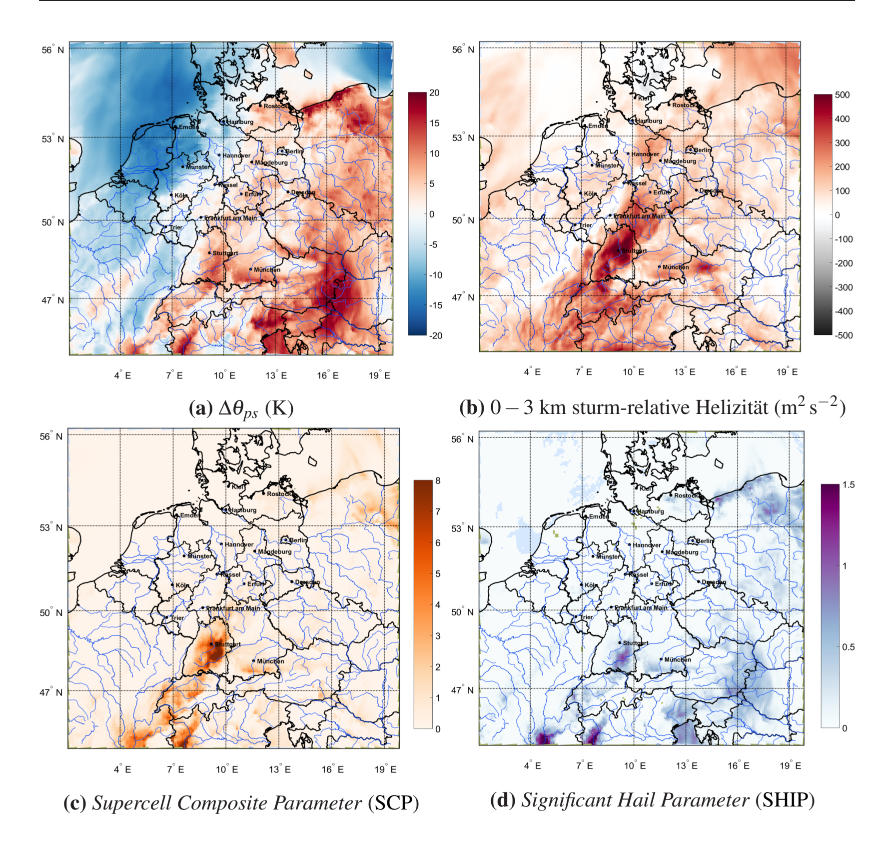
D Ergänzende Abbildungen



(a) 500 hPa Geopotential ($m^2\,s^{-2}$)

(b) 850 hPa pseudopotentielle Temperatur (K)

Abbildung D.1: Räumliche Verteilung des 500 hPa Geopotentials und der 850 hPa pseudopotentiellen Temperatur über dem COSMO-EU-Gebiet, berechnet aus der Assimilationsanalyse von COSMO-EU für den 28. Juli 2013 (15 UTC).
Gegen 13 UTC hatten sich zwei Superzellen über dem Schwarzwald gebildet, die nördlich der Schwäbischen Alb entlangzogen und Schäden in Milliardenhöhe verursachten (Kunz et al., 2018). Deutschland lag an diesem Tag in einer südwestlichen Höhenströmung auf der Vorderseite eines ausgeprägten Trogs über dem Ostatlantik. Eine Luftmassengrenze trennte hierbei mediterrane feucht-heiße Luft in der Südosthälfte (850 hPa $\theta_{ps} > 335$ K) von mäßig kühler Atlantikluft in der Nordwesthälfte Deutschlands (850 hPa $\theta_{ps} < 310$ K). Fortsetzung in Abbildung D.2.



(a) $\Delta\theta_{ps}$ (K)

(b) 0 – 3 km sturm-relative Helizität ($m^2\,s^{-2}$)

(c) *Supercell Composite Parameter* (SCP)

(d) *Significant Hail Parameter* (SHIP)

Abbildung D.2: Räumliche Verteilung verschiedener Umgebungsvariablen über Mitteleuropa, berechnet aus der Assimilationsanalyse von COSMO-EU für den 28. Juli 2013 (15 UTC). (Fortsetzung zu Abbildung D.1) In der feucht-heißen Luft wiesen hohe $\Delta\theta_{ps}$-Werte (Differenz der bodennahen θ_{ps} und der 300 hPa θ_{ps}) von regional über 10 K auf eine hohe potentielle Instabilität hin. Im Grenzbereich zur kühleren Luft ließ sich ein Maximum der DLS von rund 30 $m\,s^{-1}$ beobachten (nicht gezeigt). Die SRH zwischen 0 und 3 km über Grund erreichte um 15 UTC teils extrem hohe Werte von über 500 $m^2\,s^{-2}$, sodass besonders über Baden-Württemberg sehr gute Bedingungen für organisierte hochreichende Konvektion sowie für großen Hagel herrschten, wie auch die kombinierten Indizes SCP und SHIP südöstlich von Stuttgart (im unmittelbaren Vorfeld der beobachteten Superzellen) anzeigen.

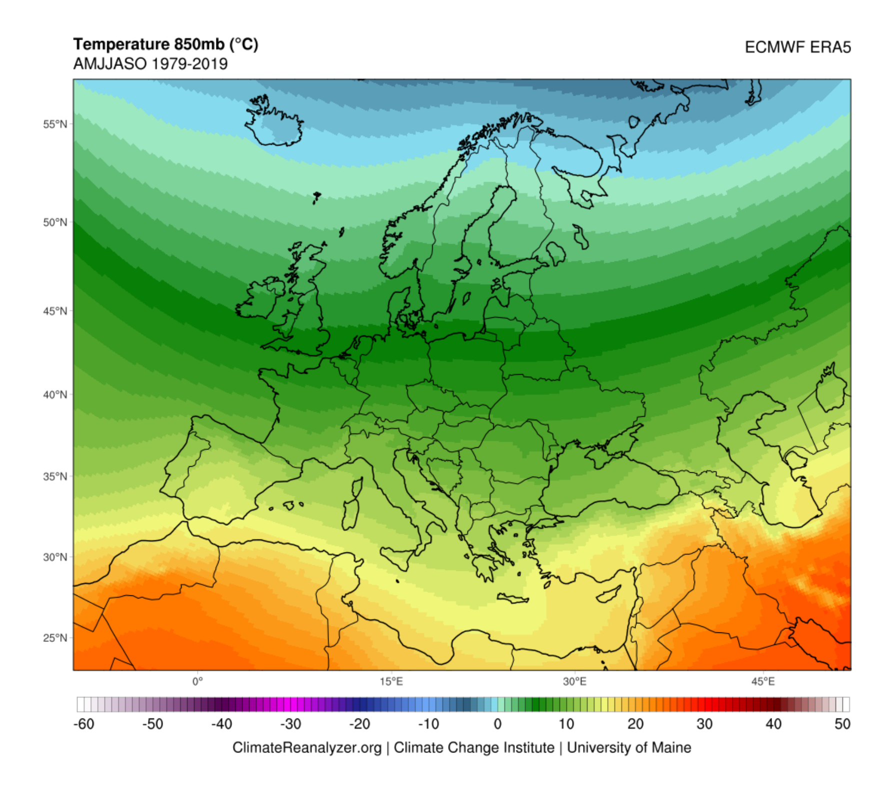
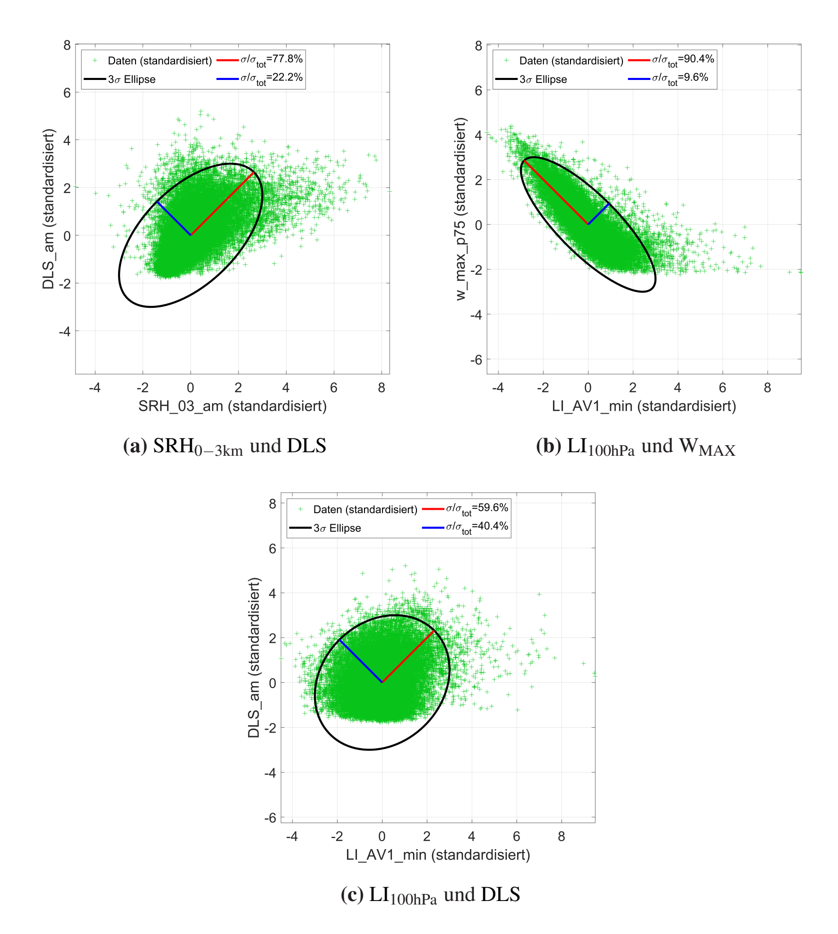


Abbildung D.3: Klimatologie der 850 hPa Temperatur (°C) über Europa für den Zeitraum 1979 – 2019, gemittelt über die Monate April bis Oktober, basierend auf Reanalysedaten (ERA5, *European Centre for Medium-Range Weather Forecasts*). Erstellt mit dem *Climate Reanalyzer* (*Climate Change Institute*, *University of Maine*, USA; `https://ClimateReanalyzer.org`). Abgerufen am 7. Januar 2021.

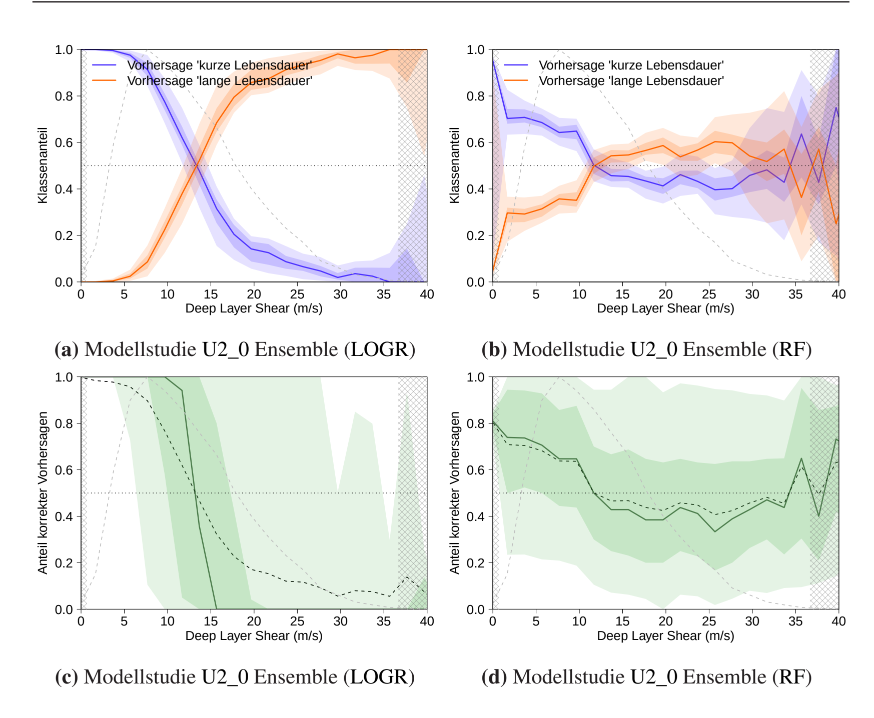


(a) SRH_{0-3km} und DLS

(b) LI_{100hPa} und W_{MAX}

(c) LI_{100hPa} und DLS

Abbildung D.4: Hauptkomponentenanalyse für drei Paare von Umgebungsvariablen zur Illustration der linearen Korrelation. Die Variablenwerte wurden zuvor standardisiert, d. h. um den Ursprung zentriert und auf den Variationsbereich $\sigma = 1$ normiert, sodass die dimensionslosen Größen vergleichbar sind (z-Transformation).



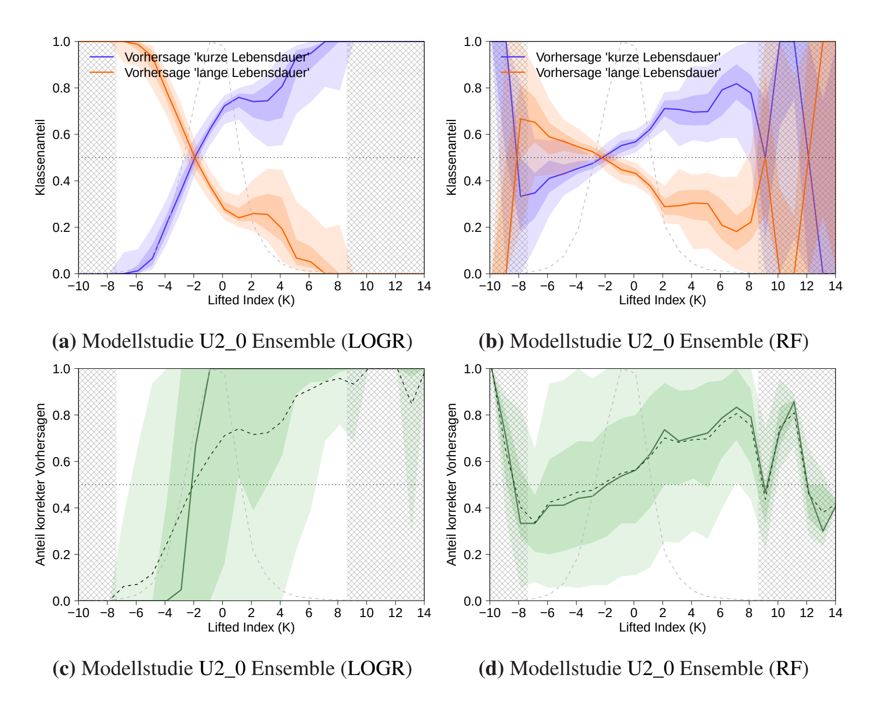
(a) Modellstudie U2_0 Ensemble (LOGR)

(b) Modellstudie U2_0 Ensemble (RF)

(c) Modellstudie U2_0 Ensemble (LOGR)

(d) Modellstudie U2_0 Ensemble (RF)

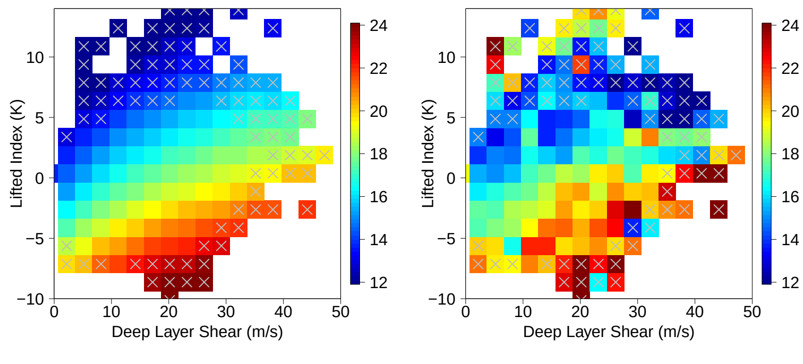
Abbildung D.5: Wie Abbildung 6.5, nur statt einer Aufteilung nach verschiedenen Werten für die Lebensdauer auf der Abszisse nun aufgespalten nach der DLS für (a,c) die logistische Regression und (b,d) den *Random Forest*.



(a) Modellstudie U2_0 Ensemble (LOGR) **(b)** Modellstudie U2_0 Ensemble (RF)

(c) Modellstudie U2_0 Ensemble (LOGR) **(d)** Modellstudie U2_0 Ensemble (RF)

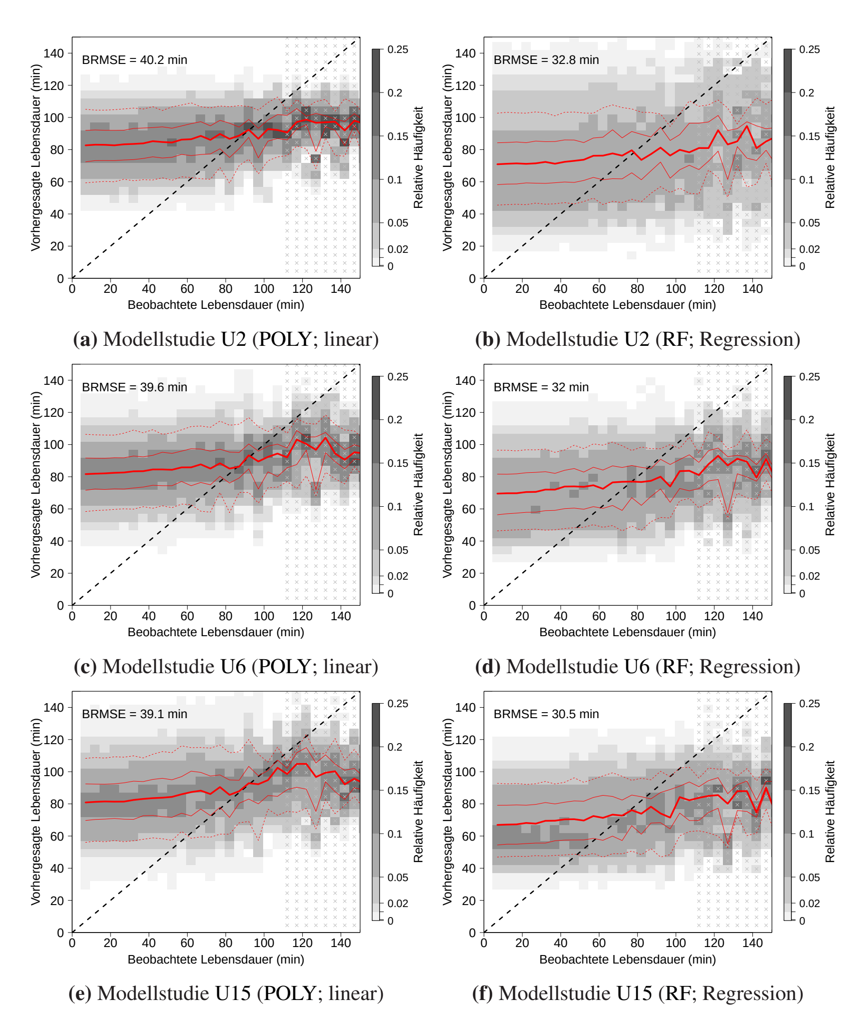
Abbildung D.6: Wie Abbildung 6.5, nur statt einer Aufteilung nach verschiedenen Werten für die Lebensdauer auf der Abszisse nun aufgespalten nach dem LI_{100hPa} für (a,c) die logistische Regression und (b,d) den *Random Forest*.



(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

Abbildung D.7: Mittlere Ensemblevorhersage für die Lebensdauer T_Z (min; Farbskala), aufgeteilt in verschiedene Gruppen der Prädiktoren DLS und $\mathrm{LI}_{100\mathrm{hPa}}$, für (a) den linearen Polynomansatz und (b) den *Random Forest* in U2 ohne *Resampling*. Graue Kreuze geben Gruppen an, in denen 20 oder weniger Zellobjekte vorliegen.



(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

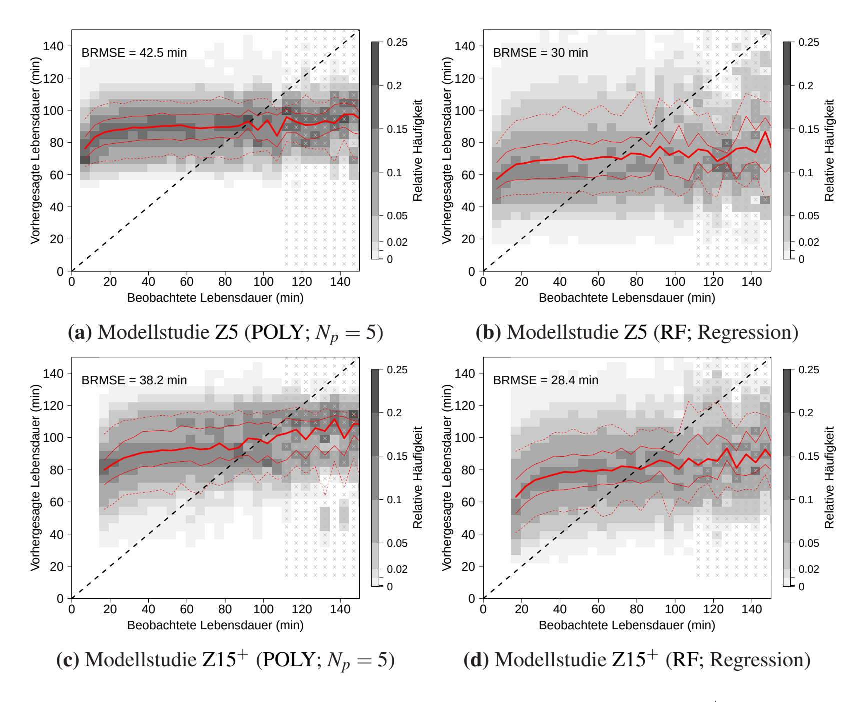
(c) Modellstudie U6 (POLY; linear)

(d) Modellstudie U6 (RF; Regression)

(e) Modellstudie U15 (POLY; linear)

(f) Modellstudie U15 (RF; Regression)

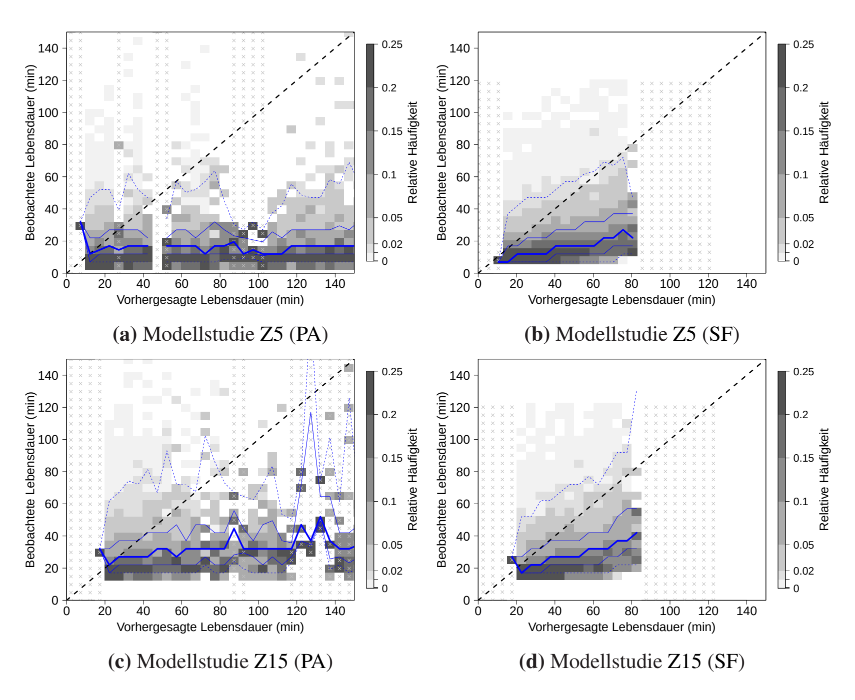
Abbildung D.8: Wie Abbildung 6.14 in (a,b) U2, (c,d) U6 und (e,f) U15. Der Polynomansatz ist hier linear.



(a) Modellstudie Z5 (POLY; $N_p = 5$) **(b)** Modellstudie Z5 (RF; Regression)

(c) Modellstudie Z15$^+$ (POLY; $N_p = 5$) **(d)** Modellstudie Z15$^+$ (RF; Regression)

Abbildung D.9: Wie Abbildung 6.14, nur in (a,b) Z5 und (c,d) Z15$^+$.



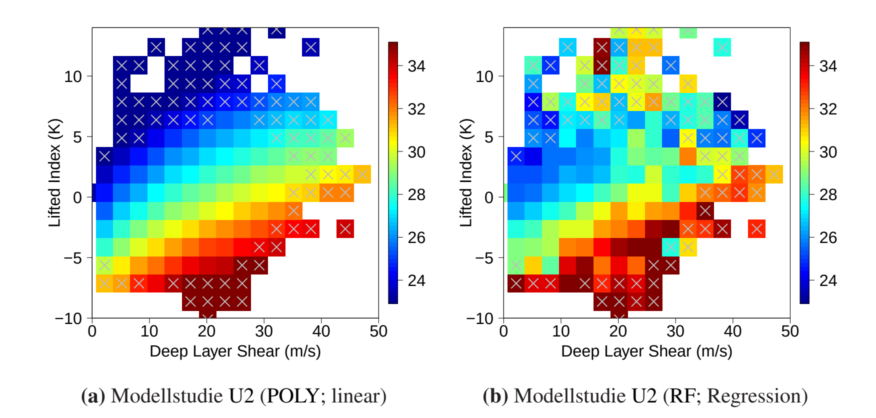
(a) Modellstudie Z5 (PA)

(b) Modellstudie Z5 (SF)

(c) Modellstudie Z15 (PA)

(d) Modellstudie Z15 (SF)

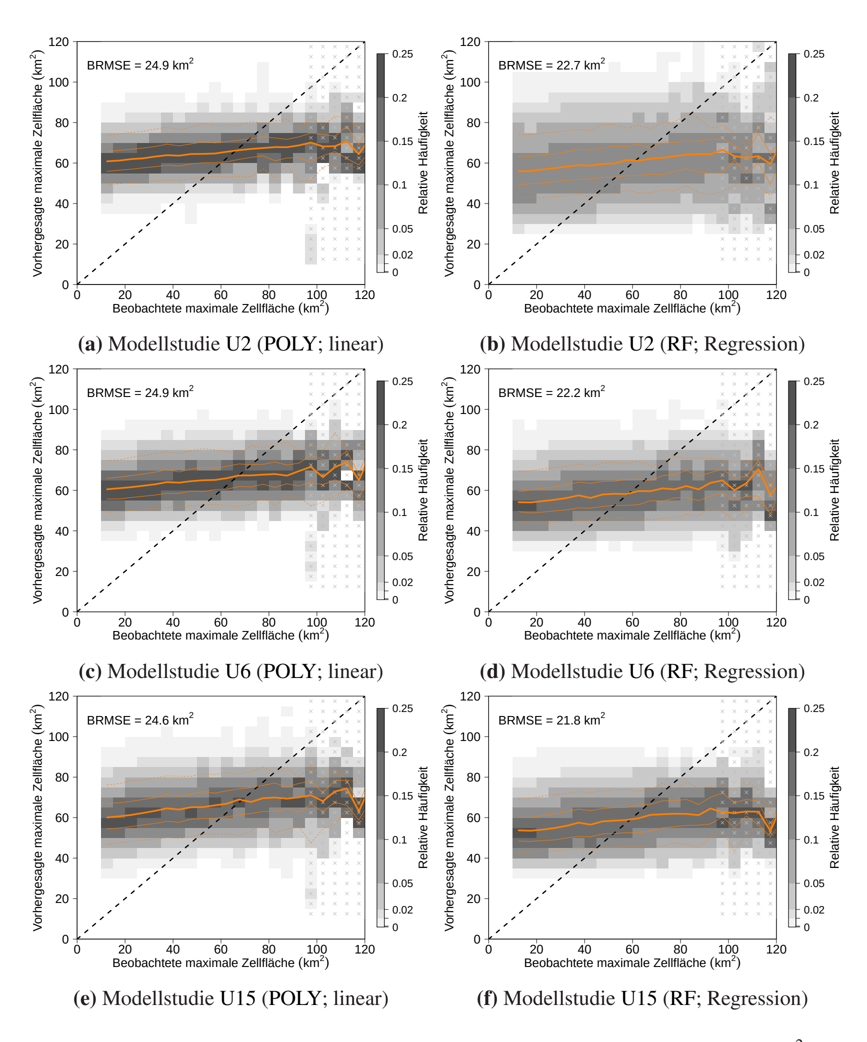
Abbildung D.10: Wie Abbildung 6.15, nur basierend auf 51 Realisierungen (a,c) des Parabelmodells und (b,d) des Modells basierend auf dem Strömungsfeldansatz. Als Prädiktor wird die Zellfläche zum Zeitpunkt der (a,b) zweiten bzw. (c,d) vierten Detektion verwendet ($A_Z(t = 7\,\text{min})$, $A_Z(t = 17\,\text{min})$).



(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

Abbildung D.11: Wie Abbildung D.7, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2; Farbskala) als Prädiktand.



(a) Modellstudie U2 (POLY; linear)

(b) Modellstudie U2 (RF; Regression)

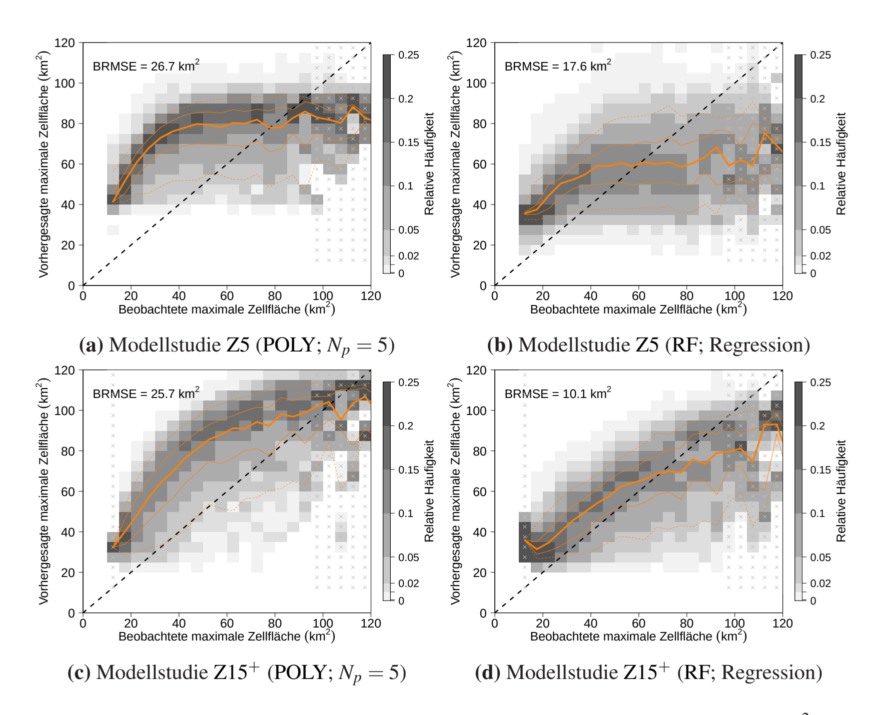
(c) Modellstudie U6 (POLY; linear)

(d) Modellstudie U6 (RF; Regression)

(e) Modellstudie U15 (POLY; linear)

(f) Modellstudie U15 (RF; Regression)

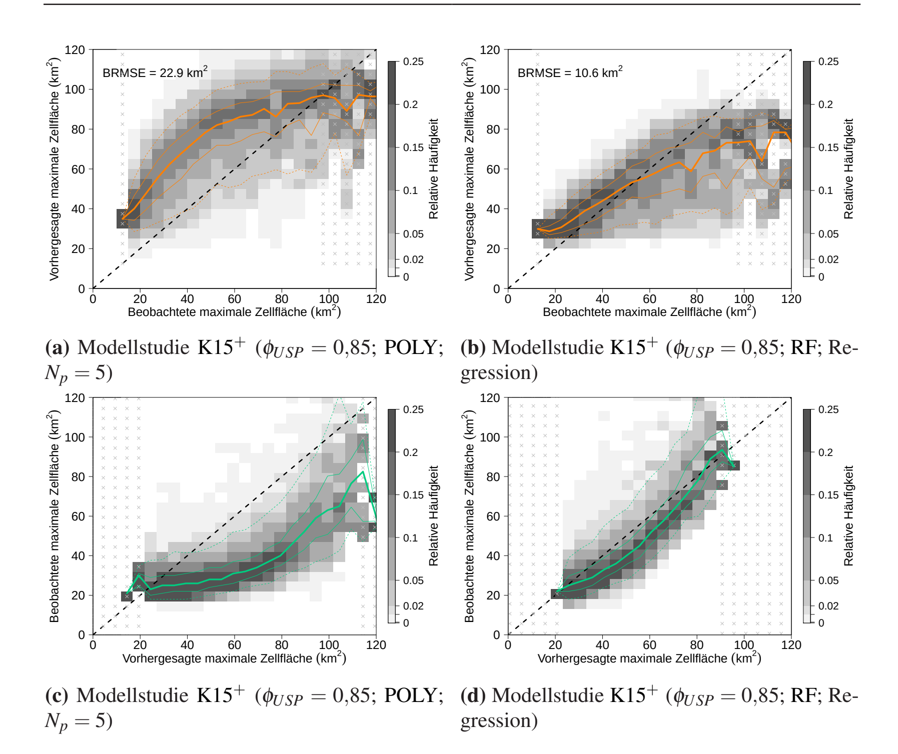
Abbildung D.12: Wie Abbildung D.8, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.



(a) Modellstudie Z5 (POLY; $N_p = 5$) **(b)** Modellstudie Z5 (RF; Regression)

(c) Modellstudie Z15$^+$ (POLY; $N_p = 5$) **(d)** Modellstudie Z15$^+$ (RF; Regression)

Abbildung D.13: Wie Abbildung D.9, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.



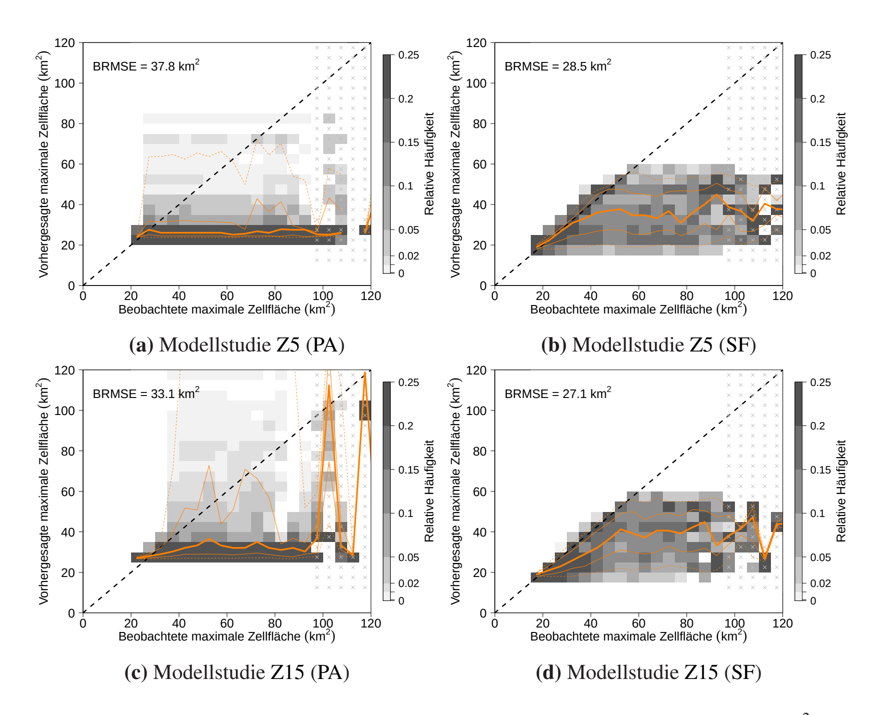
(a) Modellstudie K15$^+$ ($\phi_{USP} = 0{,}85$; POLY; $N_p = 5$)

(b) Modellstudie K15$^+$ ($\phi_{USP} = 0{,}85$; RF; Regression)

(c) Modellstudie K15$^+$ ($\phi_{USP} = 0{,}85$; POLY; $N_p = 5$)

(d) Modellstudie K15$^+$ ($\phi_{USP} = 0{,}85$; RF; Regression)

Abbildung D.14: Wie Abbildungen 6.22c+d und 6.23c+d, nur mit einem modifizierten *Resampling* mit $\phi_{USP} = 0{,}85$.



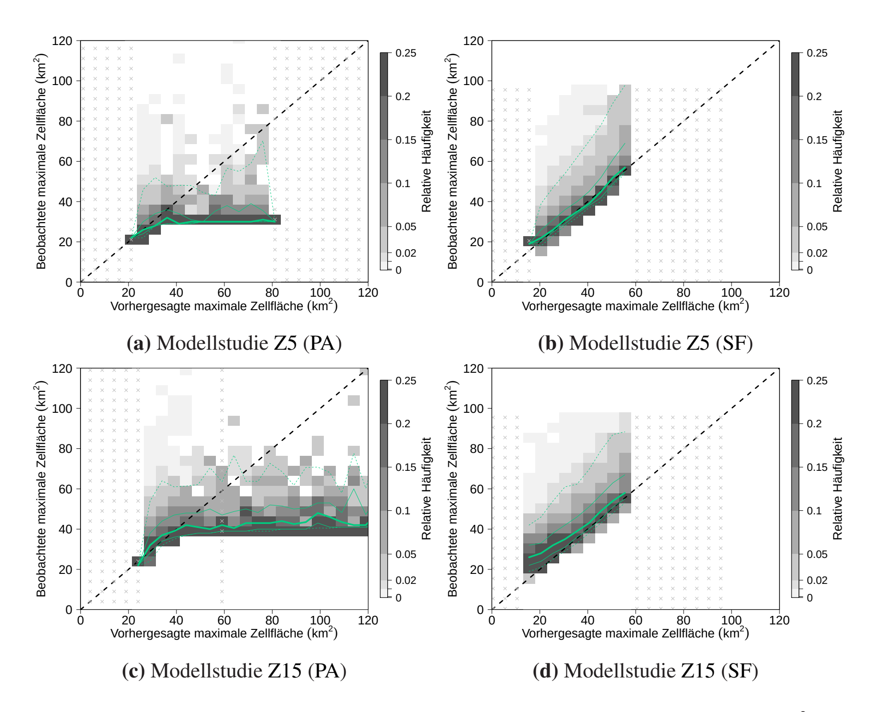
(a) Modellstudie Z5 (PA)

(b) Modellstudie Z5 (SF)

(c) Modellstudie Z15 (PA)

(d) Modellstudie Z15 (SF)

Abbildung D.15: Wie Abbildung 6.17, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.



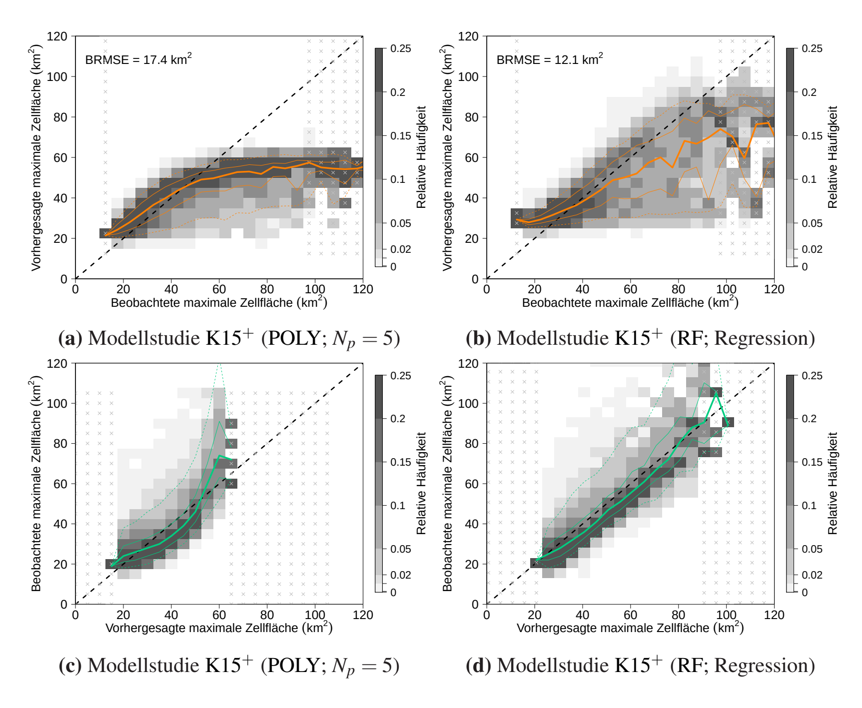
(a) Modellstudie Z5 (PA)

(b) Modellstudie Z5 (SF)

(c) Modellstudie Z15 (PA)

(d) Modellstudie Z15 (SF)

Abbildung D.16: Wie Abbildung D.10, nur mit der maximalen Zellfläche $A_{Z,max}$ (km^2) als Prädiktand.



(**a**) Modellstudie K15$^+$ (POLY; $N_p = 5$) (**b**) Modellstudie K15$^+$ (RF; Regression)

(**c**) Modellstudie K15$^+$ (POLY; $N_p = 5$) (**d**) Modellstudie K15$^+$ (RF; Regression)

Abbildung D.17: Wie Abbildungen 6.22c+d und 6.23c+d, nur ohne *Resampling*.

E Ergänzende Tabellen

Tabelle E.1: Übersicht über die verwendeten Abkürzungen von den ab Kapitel 5.2 benutzten, relevanten Umgebungsvariablen im Fließtext sowie in Abbildungen mit jeweiliger Beschreibung. Allgemeine Details zu den Variablen finden sich in den Kapiteln 2 und 4.3.3 sowie in Anhang A. Die in der Spalte Statistik angegebenen Kürzel stehen für das arithmetische Mittel in der Umgebung (`am`), das abstandsgewichtete Mittel (`wm`), die jeweiligen Standardabweichungen (`sdam`, `sdwm`), den Minimal- bzw. Maximalwert (`min`, `max`) oder Perzentilwerte (`p25`, `p50`, `p75`).

Text	Abbildung	Statistik	Beschreibung
BRN_{MU}	`BRN_MU`	`sdam`	MU *Bulk Richardson Number*
$CAPE_{MU}$	`CAPE_MU`	`wm`, `p75`	MU Konvektiv verfügbare pot. Energie
CIN_{MU}	`CIN_MU`	`p25`	MU Konvektive Hemmung
DCI_{100hPa}	`DCI_1`	`max`	ML *Deep Convective Index*
DLS	`DLS`	`wm`, `am`	*Deep Layer Shear*
$\Delta\theta_{ps}$	`DTHETAE`	`am`	Vertikaldifferenz der pseudopot. Temp. θ_{ps}
NNA_{MU}	`EL_MU`	`max`	MU Niveau des neutralen Aufstiegs
500 hPa Geop.	`FI_500`	`wm`	500 hPa Geopotential
0 °C-Grenze	`HZEROCL`	`wm`, `p25`	Höhe der Nullgradgrenze
KO-Index	`KO_INDEX`	`min`	KO-Index
HKN_{MU}	`LCL_MU`	`max`	MU Hebungskondensationsniveau
NFK_{MU}	`LFC_MU`	`min`	MU Niveau freier Konvektion
LI_{100hPa}	`LI_AV1`	`wm`, `min`	ML *Lifted Index*
$LR_{700-500hPa}$	`LR_75`	`min`	Mittlere *Lapse Rate* zw. 700 und 500 hPa

(Fortsetzung Tabelle E.1)

Text	Abbildung	Statistik	Beschreibung
$LR_{850-500hPa}$	LR_855	wm	Mittlere *Lapse Rate* zw. 850 und 500 hPa
MLS	MLS	p25	*Medium Layer Shear*
RH_{700hPa}	RH_700	min	700 hPa relative Luftfeuchtigkeit
SCP	SCP	am	*Supercell Composite Parameter*
SHIP	SHIP	max	*Significant Hail Parameter*
SI	SI	min	*Showalter Index*
SLI	SLI	wm	*Surface Lifted Index*
SRH_{0-1km}	SRH_01	max	0 – 1 km sturm-relative Helizität
$SRH_{0-1,5km}$	SRH_015	p75	0 – 1,5 km sturm-relative Helizität
SRH_{0-3km}	SRH_03	wm, am	0 – 3 km sturm-relative Helizität
T_{2m}	T_2M	min	2 m Temperatur
T_{850hPa}	T_850	wm, max, sdam	850 hPa Temperatur
$\theta_{ps,850hPa}$	THETAE_850	max, sdam	850 hPa pseudopotentielle Temperatur
IWV	TQV	wm, p25	Vertikal integrierter Wasserdampfgehalt
$\overline{U}_{0-3km}$	UMEAN_03	am	Gemittelter Horizontalwind (0 – 3 km ü. G.)
$\overline{U}_{0-6km}$	UMEAN_06	max	Gemittelter Horizontalwind (0 – 6 km ü. G.)

(Fortsetzung Tabelle E.1)

Text	Abbildung	Statistik	Beschreibung
$\overline{\mathbf{U}}_{0-10\mathrm{km}}$	`UMEAN_010`	`p50`	Gemittelter Horizontalw. (0 – 10 km ü. G.)
$\overline{\mathbf{U}}_{3-6\mathrm{km}}$	`UMEAN_36`	`am`	Gemittelter Horizontalwind (3 – 6 km ü. G.)
VT	`VERT_TOT`	`p75`	*Vertical Totals*

Tabelle E.2: Brier (Skill) Scores für die Ensemblevorhersage von 51 unterschiedlichen Modellen für die logistische Regression. Die Referenzvorhersagen stehen für konstante Vorhersagen (immer lange/kurze Lebensdauer), für die in Kapitel 3.6.1 beschriebene 50 %-Vorhersage (unsicher) und die Vorhersage, die zufällig aus einer uniformen Verteilung der Eintrittswahrscheinlichkeiten zieht (Zufall uniform), sowie eine Vorhersage, die zufällig aus der originalen Verteilung der Eintrittshäufigkeiten zieht (Zufall Verteilung).

Auswahl → Score ↓	alle Zellobjekte	nur lange Lebensdauer	nur kurze Lebensdauer
BS log. Regression	0,35	0,17	0,36
BS immer lang	0,97	0,00	1,00
BS immer kurz	0,03	1,00	0,00
BS unsicher	0,25	0,25	0,25
BS Zufall uniform	0,33	0,33	0,33
BS Zufall Verteilung	0,05	0,97	0,03
BSS immer lang	0,64	—	0,64
BSS immer kurz	−11,45	0,83	—
BSS unsicher	−0,42	0,32	−0,44
BSS Zufall uniform	−0,06	0,49	−0,08
BSS Zufall Verteilung	−5,55	0,83	−11,39

Tabelle E.3: Wie Tabelle E.2, nur für den *Random Forest* und ohne Darstellung der *BS* für die Referenzvorhersagen.

Auswahl → Score ↓	alle Zellobjekte	nur lange Lebensdauer	nur kurze Lebensdauer
BS Random Forest	0,26	0,15	0,27
BSS immer lang	0,73	—	0,73
BSS immer kurz	−8,29	0,85	—
BSS unsicher	−0,06	0,38	−0,07
BSS Zufall uniform	0,21	0,54	0,20
BSS Zufall Verteilung	−3,89	0,84	−8,21

Danksagung

Die vorliegende Arbeit entstand am Institut für Meteorologie und Klimaforschung (IMK-TRO) des Karlsruher Instituts für Technologie (KIT) im engen Austausch mit dem Deutschen Wetterdienst (DWD). Dem Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) danke ich zunächst für die Finanzierung meines Forschungsprojekts, welche diese fruchtbare Zusammenarbeit ermöglichte. Dem Deutschen Wetterdienst danke ich für die Bereitstellung der Daten des Zellverfolgungsalgorithmus KONRAD und der COSMO-Assimilationsanalysen, sowie für die Möglichkeit vor Ort in Offenbach und vom KIT aus seine HPC-Systeme zu nutzen.

Ein ganz besonderes Dankeschön gilt meinem Doktorvater Prof. Dr. Michael Kunz, der mir stets viel Vertrauen entgegenbrachte, große Freiräume zur Entwicklung und Umsetzung eigener Ideen gab und gleichzeitig bei konkreten Fragen immer für konstruktive Diskussionen zur Verfügung stand. Durch die regelmäßigen Reisen nach Offenbach, die er mir für den Austausch und die Zusammenarbeit mit dem DWD ermöglichte, entwickelte sich im Verlauf meines Forschungsprojekts ein fruchtbarer Synergismus. Die angenehme Arbeitsatmosphäre in unserer Arbeitsgruppe Atmosphärische Risiken am IMK-TRO, die Michael schafft, legte zudem den Grundstein für meine motivierte Umsetzung des Projekts. Ein besonderer Dank gilt ihm auch dafür, dass er mir die nötige Zeit und die Freiräume unter anderem für mein Engagement im Rahmen der JGW-Nachhaltigkeitsakademien 2018 und 2019 gegeben hat.

Vielen Dank an Prof. Dr. Roland Potthast und Prof. Dr. Christoph Kottmeier für die Übernahme des Korreferats, die Unterstützung des Projekts sowie wertvolle Diskussionen. Roland, dir gilt ein besonderer Dank für die spannenden Gespräche über verschiedene mathematische und numerische Ansätze, die ich sehr genossen habe. Mein Dank geht zudem an Dr. Ulrich Blahak, Dr. Robert Feger und Dr. Kathrin Wapler, die mich und das Projekt seitens des DWD begleitet haben. Der Austausch bei unseren Treffen war sehr anregend, sodass wir zusammen viele Ideen entwickeln konnten. Außerdem vielen Dank für eure Hilfe bei der Beseitigung der ein oder anderen technischen Schwierigkeit. Insbesondere dir, Uli, herzlichen Dank für dein Engagement und die Anbindung an SINFONY, und danke allen SINFONY-Mitarbeitern, die ich in den letzten Jahren kennen lernen durfte, für die freundliche Aufnahme in das Omega-Haus bei meinen Reisen nach Offenbach und die gemütlichen Abende während der ECSS in Krakau.
Ein riesiges Dankeschön geht außerdem an die gesamte Arbeitsgruppe Atmosphärische Risiken. Egal ob auf der Arbeit, beim Skifahren, Bogenschießen oder Grillen, die Atmosphäre ist stets positiv und freundschaftlich und führt dazu, dass ich jeden Tag gerne ins Büro komme – sofern die äußeren Umstände dies erlauben, die uns seit vielen Monaten mehr oder weniger an unser Heimbüro fesseln. Danke also an Sanna, Heinz Jürgen, Melanie, Manu, Markus, Sinan und Mathis für die schönen letzten drei Jahre und vielleicht ja noch das ein oder andere zukünftige gemeinsame Jahr. Manu, dir vielen Dank für diverse fachliche und nicht-fachliche Gespräche über das Wetter und die Welt. Melanie und Heinz Jürgen, euch vielen Dank für die gemeinsamen sportlichen Aktivitäten (Ob wir jemals einen Kraulschwimmkurs zu Ende führen dürfen?). Sanna, dir vielen Dank für deine stete Hilfsbereitschaft, den regen Austausch zwischen unseren benachbarten Büros (auch von Tee, Schaumküssen oder anderen Leckereien) und dein großes Einfühlungsvermögen. Natürlich auch ein großes Dankeschön für deine wertvollen Kommentare während des Schreibprozesses dieser Dissertation, welches ich ebenso an Flo richte, der sich ebenfalls mühevoll durch viele Seiten hindurchgewälzt und hilfreiche

Hinweise gegeben hat. Flo, dir ohnehin ein kräftiges Danke für unseren guten Austausch und die vielen musikalisch-fröhlichen Abendstunden, die wir mit unseren Freunden beim KIT-Konzertchor bislang verbracht haben.

Vielen Dank auch für die große Unterstützung durch meine weiteren Kollegen am IMK-TRO. Ich danke Uli Corsmeier und Roswitha Marioth für einen wohlwollenden Rahmen, Jan Handwerker für spannende Tage in den Tiefen der Radarmeteorologie, Joaquim Pinto und Patrick Ludwig für ein interessantes IPCC-Seminar sowie Peter Knippertz, Corinna Hoose und Christian Grams für den angenehmen Austausch während meiner Übungsgruppenleitung zu den Theorie-Vorlesungen. Ein besonderer Dank gilt Hans Schipper für das gute Zusammenwirken bei unseren gemeinsamen *Outreach*-Aktivitäten. Hans, danke auch für die vielen spannenden Klimawandel-Diskussionen und gedanklichen Ausflüge in die Niederlande während des Mittagessens. Heike Vogel danke ich für den warmen Empfang am IMK-TRO und die freundliche Hilfsbereitschaft zur Beseitigung technischer COSMO-Probleme. Ein großer Dank geht zudem an Gabi Klinck und Gerhard Brückel für eine hervorragende Administration der Infrastruktur und die stete Hilfsbereitschaft bei technischen Fragen. Gabi, danke für die vielen aufmunternden Gespräche zu späterer Stunde am Institut insbesondere an Tagen, an denen nicht alles so läuft, wie man es sich wünscht. Vielen Dank auch an Doris und Rosi für das stets zuverlässige und angenehme Regeln von organisatorischen Dingen aller Art. Auch allen übrigen Kollegen, die ich in den letzten Jahren kennen lernen durfte, danke ich sehr herzlich für die angenehme Atmosphäre.

Ein besonders großer Dank geht natürlich an meine ehemaligen Kommilitonen, meine Freunde und meine Familie. Ich danke meiner Oma Erika, meiner Tante Christel sowie meinen Paten Elfi und Rainer, die sowohl meine musikalische Jugend als auch mein wissenschaftliches Studium stets mit Interesse verfolgt und unterstützt haben. Nicht zu vergessen natürlich der Rest meiner wunderbaren, herzlichen saarländischen Großfamilie! Tino, du bist jeden Tag an meiner Seite und unterstützt mich bei allem, was ich tue.

Du hast mir den Rücken an so vielen Tagen freigehalten, an denen ich mich voll und ganz auf die Dissertation konzentrieren konnte. Wenn es nötig ist, munterst du mich auf. Jeden Tag aufs Neue schenkst du mir Vertrauen und Kraft. Dafür bin ich dir unendlich dankbar. Anna, du bist nicht nur eine hervorragende Korrekturleserin, sondern auch eine wunderbare Schwester, die immer einen besonderen Platz in meinem Leben haben wird. Zuletzt gilt ein großes, herzliches Dankeschön meinen Eltern Ingrid und Christian, die zu jeder Zeit für mich da sind (auch wenn uns knapp 200 Kilometer trennen) und mir nicht nur während des Studiums und der Promotion viel Kraft gegeben haben. Danke, dass es euch gibt!

Karlsruhe, im Januar 2021 *Jannik Wilhelm*

Wissenschaftliche Berichte des Instituts für Meteorologie und Klimaforschung des Karlsruher Instituts für Technologie (0179-5619)

Bisher erschienen:

Nr. 1: *Fiedler, F. / Prenosil, T.*
Das MESOKLIP-Experiment. (Mesoskaliges Klimaprogramm im Oberrheintal)

Nr. 2: *Tangermann-Dlugi, G.*
Numerische Simulationen atmosphärischer Grenzschichtströmungen über langgestreckten mesoskaligen Hügelketten bei neutraler thermischer Schichtung

Nr. 3: *Witte, N.*
Ein numerisches Modell des Wärmehaushalts fließender Gewässer unter Berücksichtigung thermischer Eingriffe

Nr. 4: *Fiedler, F. / Höschele, K. (Hrsg.)*
Prof. Dr. Max Diem zum 70. Geburtstag

Nr. 5: *Adrian, G.*
Ein Initialisierungsverfahren für numerische mesoskalige Strömungs modelle

Nr. 6: *Dorwarth, G.*
Numerische Berechnung des Druckiderstandes typischer Geländeformen

Nr. 7: *Vogel, B.; Adrian, G. / Fiedler, F.*
MESOKLIP-Analysen der meteorologischen Beobachtungen von mesoskaligen Phänomenen im Oberrheingraben

Nr. 8: *Hugelmann, C.-P.*
Differenzenverfahren zur Behandlung der Advektion

Nr. 9: *Hafner, T.*
Experimentelle Untersuchung zum Druckwiderstand der Alpen

Nr. 10: *Corsmeier, U.*
Analyse turbulenter Bewegungsvorgänge in der maritimen atmosphärischen Grenzschicht

Nr. 11: *Walk, O. / Wieringa, J.(eds)*
Tsumeb Studies of the Tropical Boundary-Layer Climate

Nr. 12: *Degrazia, G. A.*
Anwendung von Ähnlichkeitsverfahren auf die turbulente Diffusion in der konvektiven und stabilen Grenzschicht

Nr. 13: *Schädler, G.*
Numerische Simulationen zur Wechselwirkung zwischen Landoberflächen und atmophärischer Grenzschicht

Nr. 14: *Heldt, K.*
Untersuchungen zur Überströmung eines mikroskaligen Hindernisses in der Atmosphäre

Nr. 15: *Vogel, H.*
Verteilungen reaktiver Luftbeimengungen im Lee einer Stadt – Numerische Untersuchungen der relevanten Prozesse

Nr. 16: *Höschele, K.(ed.)*
Planning Applications of Urban and Building Climatology – Proceedings of the IFHP / CIB-Symposium Berlin, October 14-15, 1991

Nr. 17: *Frank, H. P.*
Grenzschichtstruktur in Fronten

Nr. 18: *Müller, A.*
Parallelisierung numerischer Verfahren zur Beschreibung von Ausbreitungs- und chemischen Umwandlungsprozessen in der atmosphärischen Grenzschicht

Nr. 19: *Lenz, C.-J.*
Energieumsetzungen an der Erdoberfläche in gegliedertem Gelände

Nr. 20: *Schwartz, A.*
Numerische Simulationen zur Massenbilanz chemisch reaktiver Substanzen im mesoskaligen Bereich

Nr. 21: *Beheng, K. D.*
Professor Dr. Franz Fiedler zum 60. Geburtstag

Nr. 22: *Niemann, V.*
Numerische Simulation turbulenter Scherströmungen mit einem Kaskadenmodell

Nr. 23: *Koßmann, M.*
Einfluß orographisch induzierter Transportprozesse auf die Struktur der atmosphärischen Grenzschicht und die Verteilung von Spurengasen

Nr. 24: *Baldauf, M.*
Die effektive Rauhigkeit über komplexem Gelände – Ein Störungstheoretischer Ansatz

Nr. 25: *Noppel, H.*
Untersuchung des vertikalen Wärmetransports durch die Hangwindzirkulation auf regionaler Skala

Nr. 26: *Kuntze, K.*
Vertikaler Austausch und chemische Umwandlung von Spurenstoffen über topographisch gegliedertem Gelände

Nr. 27: *Wilms-Grabe, W.*
Vierdimensionale Datenassimilation als Methode zur Kopplung zweier verschiedenskaliger meteorologischer Modellsysteme

Nr. 28: *Grabe, F.*
Simulation der Wechselwirkung zwischen Atmosphäre, Vegetation und Erdoberfläche bei Verwendung unterschiedlicher Parametrisierungsansätze

Nr. 29: *Riemer, N.*
Numerische Simulationen zur Wirkung des Aerosols auf die troposphärische Chemie und die Sichtweite

Nr. 30: *Braun, F. J.*
Mesoskalige Modellierung der Bodenhydrologie

Nr. 31: *Kunz, M.*
Simulation von Starkniederschlägen mit langer Andauer über Mittelgebirgen

Nr. 32: *Bäumer, D.*
Transport und chemische Umwandlung von Luftschadstoffen im Nahbereich von Autobahnen – numerische Simulationen

Nr. 33: *Barthlott, C.*
Kohärente Wirbelstrukturen in der atmosphärischen Grenzschicht

Nr. 34: *Wieser, A.*
Messung turbulenter Spurengasflüsse vom Flugzeug aus

Nr. 35: *Blahak, U.*
Analyse des Extinktionseffektes bei Niederschlagsmessungen mit einem C-Band Radar anhand von Simulation und Messung

Nr. 36: *Bertram, I.*
Bestimmung der Wasser- und Eismasse hochreichender konvektiver Wolken anhand von Radardaten, Modellergebnissen und konzeptioneller Betrachtungen

Nr. 37: *Schmoeckel, J.*
Orographischer Einfluss auf die Strömung abgeleitet aus Sturmschäden im Schwarzwald während des Orkans „Lothar“

Nr. 38: *Schmitt, C.*
Interannual Variability in Antarctic Sea Ice Motion: Interannuelle Variabilität antarktischer Meereis-Drift

Nr. 39: *Hasel, M.*
Strukturmerkmale und Modelldarstellung der Konvektion über Mittelgebirgen

Ab Band 40 erscheinen die Wissenschaftlichen Berichte des Instituts für Meteorologie und Klimaforschung bei KIT Scientific Publishing (ISSN 0179-5619). Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Nr. 40: *Lux, R.*
Modellsimulationen zur Strömungsverstärkung von orographischen Grundstrukturen bei Sturmsituationen
ISBN 978-3-86644-140-8

Nr. 41: *Straub, W.*
Der Einfluss von Gebirgswellen auf die Initiierung und Entwicklung konvektiver Wolken
ISBN 978-3-86644-226-9

Nr. 42: *Meißner, C.*
High-resolution sensitivity studies with the regional climate model COSMO-CLM
ISBN 978-3-86644-228-3

Nr. 43: *Höpfner, M.*
Charakterisierung polarer stratosphärischer Wolken mittels hochauflösender Infrarotspektroskopie
ISBN 978-3-86644-294-8

Nr. 44: *Rings, J.*
Monitoring the water content evolution of dikes
ISBN 978-3-86644-321-1

Nr. 45: *Riemer, M.*
Außertropische Umwandlung tropischer Wirbelstürme: Einfluss auf das Strömungsmuster in den mittleren Breiten
ISBN 978-3-86644-766-0

Nr. 46: Nicht erschienen

Nr. 47: Nicht erschienen

Nr. 48: Nicht erschienen

Nr. 49: *Peters, T.*
Ableitung einer Beziehung zwischen der Radarreflektivität, der Niederschlagsrate und weiteren aus Radardaten abgeleiteten Parametern unter Verwendung von Methoden der multivariaten Statistik
ISBN 978-3-86644-323-5

Nr. 50: *Khodayar Pardo, S.*
High-resolution analysis of the initiation of deep convection forced by boundary-layer processes
ISBN 978-3-86644-770-7

Nr. 51: *Träumner, K.*
Einmischprozesse am Oberrand der konvektiven atmosphärischen Grenzschicht
ISBN 978-3-86644-771-4

Nr. 52: Nicht erschienen

Nr. 53: *Lundgren, K.*
Direct Radiative Effects of Sea Salt on the Regional Scale
ISBN 978-3-86644-773-8

Nr. 54: *Sasse, R.*
Analyse des regionalen atmosphärischen Wasserhaushalts unter Verwendung von COSMO-Simulationen und GPS-Beobachtungen
ISBN 978-3-86644-774-5

Nr. 55: *Grenzhäuser, J.*
Entwicklung neuartiger Mess- und Auswertungsstrategien für ein scannendes Wolkenradar und deren Anwendungsbereiche
ISBN 978-3-86644-775-2

Nr. 56: *Grams, C.*
Quantification of the downstream impact of extratropical transition for Typhoon Jangmi and other case studies
ISBN 978-3-86644-776-9

Nr. 57: *Keller, J.*
Diagnosing the Downstream Impact of Extratropical Transition Using Multimodel Operational Ensemble Prediction Systems
ISBN 978-3-86644-984-8

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Nr. 58: *Mohr, S.*
Änderung des Gewitter- und Hagelpotentials im Klimawandel
ISBN 978-3-86644-994-7

Nr. 59: *Puskeiler, M.*
Radarbasierte Analyse der Hagelgefährdung in Deutschland
ISBN 978-3-7315-0028-5

Nr. 60: *Zeng, Y.*
Efficient Radar Forward Operator for Operational Data Assimilation within the COSMO-model
ISBN 978-3-7315-0128-2

Nr. 61: Nicht erschienen

Nr. 62: *Jerger, D.*
Radar Forward Operator for Verification of Cloud Resolving Simulations within the COSMO Model
ISBN 978-3-7315-0172-5

Nr. 63: *Maurer, V.*
Vorhersagbarkeit konvektiver Niederschläge: Hochauflösende Ensemblesimulationen für Westafrika
ISBN 978-3-7315-0189-3

Nr. 64: *Stawiarski, C.*
Optimizing Dual-Doppler Lidar Measurements of Surface Layer Coherent Structures with Large-Eddy Simulations
ISBN 978-3-7315-0197-8

Nr. 65: *Mahlke, H.*
Mechanismen der Auslösung hochreichender Konvektion im südwestdeutschen Mittelgebirgsraum
ISBN 978-3-7315-0203-6

Nr. 66: *Fosser, G.*
Precipitation statistics from regional climate model at resolutions relevant for soil erosion
ISBN 978-3-7315-0227-2

Nr. 67: *Adler, B.*
Boundary-Layer Processes Producing Mesoscale Water-Vapour Variability over a Mountainous Island
ISBN 978-3-7315-0247-0

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Nr. 68: *Kraut, I.*
Separating the Aerosol Effect in Case of a „Medicane"
ISBN 978-3-7315-0405-4

Nr. 69: *Breil, M.*
Einfluss der Boden-Vegetation-Atmosphären Wechselwirkungen auf die dekadische Vorhersagbarkeit des Westafrikanischen Monsuns
ISBN 978-3-7315-0420-7

Nr. 70: *Lott, F. F.*
Wind Systems in the Dead Sea and Footprints in Seismic Records
ISBN 978-3-7315-0596-9

Nr. 71: *Rieger, D.*
Der Einfluss von natürlichem Aerosol auf Wolken über Mitteleuropa
ISBN 978-3-7315-0672-0

Nr. 72: *Loewe, K.*
Arctic mixed-phase clouds. Macro- and microphysical insights with a numerical model
ISBN 978-3-7315-0686-7

Nr. 73: *Piper, D. A.*
Untersuchung der Gewitteraktivität und der relevanten großräumigen Steuerungsmechanismen über Mittel- und Westeuropa
ISBN 978-3-7315-0701-7

Nr. 74: *Metzger, J.*
Wind Systems and Energy Balance in the Dead Sea Valley
ISBN 978-3-7315-0699-7

Nr. 75: *Deetz, K.*
Assessing the Aerosol Impact on Southern West African Clouds and Atmospheric Dynamics
ISBN 978-3-7315-0744-4

Nr. 76: *Ehmele, F. T.*
Stochastische Simulation großflächiger, hochwasserrelevanter Niederschlagsereignisse
ISBN 978-3-7315-0761-1

Nr. 77: *Hackenbruch, J.*
Anpassungsrelevante Klimaänderungen für städtische Baustrukturen und Wohnquartiere
ISBN 978-3-7315-0771-0

Nr. 78: *Schmidberger, M.*
Hagelgefährdung und Hagelrisiko in Deutschland basierend auf einer Kombination von Radardaten und Versicherungsdaten
ISBN 978-3-7315-0846-5

Nr. 79: *Gruber, S.*
Contrails and Climate Engineering - Process Studies on Natural and Artificial High-Level Clouds and Their Impact on the Radiative Fluxes
ISBN 978-3-7315-0896-0

Nr. 80: *Walter, C.*
Simulationen der Ausbreitung von Vulkanasche unter expliziter Berücksichtigung der optischen Eigenschaften der Aschepartikel
ISBN 978-3-7315-0939-4

Nr. 81: *Caldas-Alvarez, A.*
Atmospheric Moisture Effects on Deep Convection in the Western Mediterranean
ISBN 978-3-7315-0947-9

Nr. 82: *Sedlmeier, K.*
Near future changes of compound extreme events from an ensemble of regional climate simulations
ISBN 978-3-7315-0476-4

Nr. 83: *Brecht, B. M.*
Die urbane Wärmebelastung unter Einfluss lokaler Faktoren und zukünftiger Klimaänderungen
ISBN 978-3-7315-0990-5

Nr. 84: *Singh, S.*
Convective precipitation simulated with ICON over heterogeneous surfaces in dependence on model and land-surface resolution
ISBN 978-3-7315-1068-0

Nr. 85: *Wilhelm, J.*
Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage
ISBN 978-3-7315-1182-3

85

JANNIK WILHELM

Einfluss atmosphärischer Umgebungsbedingungen auf den Lebenszyklus konvektiver Zellen in der Echtzeit-Vorhersage

Präzise Warnungen vor Gewittern und ihren Begleiterscheinungen wie Starkregen, Hagel oder Sturmböen sind für kurzfristige Präventionsmaßnahmen bis hin zur Evakuierung von Menschen unerlässlich. Die dynamische Entwicklung von konvektiven Zellen während instabiler Wetterlagen führt oftmals zu einer erheblichen Diskrepanz zwischen den Echtzeit-Vorhersagen der Wetterdienste und den beobachteten Wetterbedingungen. Daten des Zellverfolgungsalgorithmus KONRAD des Deutschen Wetterdienstes werden mit COSMO-EU Modellanalysen kombiniert, um den Lebenszyklus von Gewittern in Deutschland und den Einfluss der atmosphärischen Umgebungsbedingungen zu analysieren sowie Verfahren aus der Statistik und dem Maschinellen Lernen zur Vorhersage der Zellgröße und Lebensdauer zu entwickeln und zu evaluieren.

ISSN 0179-5619
ISBN 978-3-7315-1182-3

Gedruckt auf FSC-zertifiziertem Papier

ISBN 978-3-7315-1182-3
9 783731 511823 >